Речевые технологии

2/2009

Главный редактор ***Александр Харламов***

Состав редколлегии:

Потапова Р.К., *доктор филологических наук, профессор, заместитель главного редактора*
Аграновский А.В., *доктор технических наук, профессор*
Женило В.Р., *доктор технических наук*
Жигулёвцев Ю.Н., *кандидат технических наук*
Кривнова О.Ф., *доктор филологических наук*
Кушнир А.М., *кандидат психологических наук*
Лобанов Б.М., *доктор технических наук (Беларусь)*
Максимов Е.М., *доктор технических наук*
Малеев О.Г., *кандидат технических наук*
Михайлов В.Г., *доктор филологических наук*
Нариньяни А.С., *кандидат физико-математических наук*
Петровский А.А., *доктор технических наук (Беларусь)*
Хитров М.В., *кандидат технических наук*
Чучупал В.Я., *кандидат физико-математических наук*
Шелепов В.Ю., *доктор физико-математических наук (Украина)*
Кушнир Д.А., *ответственный секретарь, кандидат технических наук*

Содержание

Редакция:
Редактор — *Артём Ганькин*
Корректор — *Ирина Дёмина*
Дизайн — *Анна Ладанюк*
Вёрстка — *Сергей Бурукин*

Адрес редакции: 109341, Москва, ул. Люблинская, д. 157, корп. 2.
Тел.: (495) 979-54-27

Подписано в печать 24.12.2009. Формат 60×90⅛. Бумага офсетная. Печать офсетная.
Печ. л. 14,25. Заказ № 0204. Издательский дом «Народное образование».
Отпечатано в типографии НИИ школьных технологий. 143500, г. Истра-2, ул. Заводская, д. 2А.
Тел.: 8 901 519-53-96, (495) 792-59-62.

© «Народное образование»

Моделирование подсвязочной области в частотно-временном артикуляторном синтезаторе

Горбунов К.С.,

Макаров И.С.,
кандидат технических наук,
старший научный сотрудник

Статья посвящена моделированию влияния подсвязочной области на акустические характеристики речевого тракта в частотно-временных артикуляторных синтезаторах. Построенная модель включает в себя вычисление кратковременного спектра Фурье голосового источника, учёт взаимодействия передаточной функции речевого тракта и подсвязочных полостей в частотной области и дальнейший синтез речевого сигнала методом наложения с суммированием. Схема протестирована на результатах измерений динамики речевого тракта методом магнитно-резонансной томографии для ряда звукосочетаний американского английского языка.

1. Постановка проблемы

Подсвязочная область представляет из себя систему полостей, расположенных ниже голосовых складок, — трахею, бронхи и лёгкие. Её влияние на акустические и аэродинамические процессы в речевом тракте разнообразно. В частности, при раскрытии голосовой щели создаются условия для возникновения дополнительных резонансов и антирезонансов, обусловленных влиянием подсвязочной области [Сорокин, 1985]. При этом резонансы речевого тракта модулируются как по амплитуде, так и по частоте. В некоторых случаях возможна бифуркация частоты первого резонанса речевого тракта, наблюдаемая на сонограммах в виде скачка амплитуды 100–300 Гц [Stevens, 2000; Chi, Sonderegger, 2007].

Таким образом, в результате взаимодействия подсвязочной области и речевого тракта в акустическом сигнале на интервале открытой голосовой щели наблюдаются спектральные компоненты, которые отсутствуют при закрытой голосовой щели. Влияние этих компонент на восприятие речи представляет значительный интерес.

По мнению [Сорокин, 1992], они определяют тембровые характеристики голоса и создают индивидуальные особенности звучания в зависимости от геометрических размеров трахеи, бронхов и лёгких. Согласно [Assmann, Katz, 2005], дополнительные резонансы и антирезонансы, порождённые подсвязочной областью, определяют не только натуральность, но и разборчивость стационарных гласных. Параметры трахеи и лёгких, вероятно, играют важную роль в создании ряда акустических эффектов в пении [Морозов, 2002].

Несмотря на отмеченные публикации, взаимодействие подсвязочной области и речевого тракта продолжает оставаться малоизученной проблемой. Нам, например, неизвестны работы, посвящённые моделированию этого взаимодействия для оценки его роли в восприятии речи.

На наш взгляд, это обстоятельство связано с трудностями моделирования подсвязочной области в существующих типах артикуляторных синтезаторов. В самом деле, во временных артикуляторных синтезаторах — то есть в синтезаторах, вычисляющих речевую волну во временной области либо методом бегущих волн, либо методом конечных разностей [Сорокин, 1992], — влияние подсвязочной области не может быть адекватно учтено из-за наличия нелинейных функций частоты в переносном импедансе трахеи. В частотно-временных синтезаторах источники возбуждения звука вычисляются во временной области, а характеристики тракта — в частотной [Сорокин, 1992]. При этом предполагается, что характеристики тракта (его передаточная функция и переносной акустический импеданс) изменяются во времени гораздо медленнее, чем площадь голосовой щели. По этой причине быстрые изменения спектральных характеристик сигнала, происходящие внутри одного периода основного тона и порождённые взаимодействием речевого тракта и подсвязочной области, принципиально не могут быть реализованы данным типом синтезатора без существенной переделки его общей схемы.

Основная задача данной работы состоит в построении такой схемы частотно-временного артикуляторного синтезатора речи, которая позволила бы моделировать влияние подсвязочной области на акустические характеристики речевого тракта. Основная идея построения заключается в вычислении кратковременного спектра Фурье голосового источника в скользящем окне анализа, учёте взаимодействия речевого тракта и подсвязочных полостей в частотной области и дальнейшем синтезе речевого сигнала методом наложения с суммированием. Предлагаемая схема напоминает процедуру ресинтеза сигнала в вокодере STRAIGHT [Kawahara et al., 1999].

Структура работы выглядит следующим образом. В разделе 2 описывается базовая схема используемого частотно-временного артикуляторного синтезатора. Раздел 3 посвящён модели взаимодействия подсвязочной области и речевого тракта. В разделе 4 строится новая схема синтезатора, учитывающая влияние трахеи, бронхов и лёгких на характеристики тракта. Раздел 5 описывает эксперименты по синтезу некоторых звукосочетаний с помощью построенной схемы. Наконец, раздел 6 содержит обсуждение полученных результатов.

2. Частотно-временной артикуляторный синтезатор речи

Частотно-временной синтезатор, используемый в данной статье, содержит несколько блоков (рис. 1).

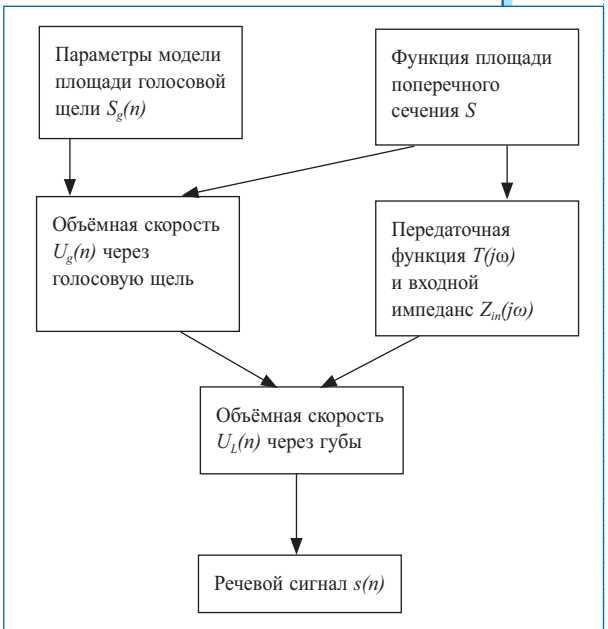


Рис. 1. Структура используемого частотно-временного артикуляторного синтезатора

Синтезатор работает на частоте дискретизации 16 кГц с квантованием каждого значения амплитуды на 16 бит. На вход синтезатора через одинаковые интервалы времени Δ (обычно полагаемые равными 10 мс) подаются значения площади поперечного сечения S и длины речевого тракта, а также значения некоторых параметров управления моделью площади голосовой щели $S_g(n)$ (n — номер дискретного отсчёта времени), включая контур основного тона. При этом значения площади тракта могут быть либо вычислены по некоторой математической модели артикуляции (например, с помощью алгоритма из [Баден и др., 2005]), либо получены путём непосредственного измерения с помощью ядерно-магнитной томографии [Story et al., 1998]. В качестве модели площади голосовой щели используется модель из [Сорокин, Макаров, 2008].

Затем по функции площади речевого тракта определяется его передаточная функция $T(j\omega)$ и некоторые импедансы (в частности, входной акустический импеданс $Z_{in}(j\omega)$ в речевой тракт со стороны голосовой щели) с помощью модифицированного метода длинной линии [Макаров, 2009]. Кроме того, на этом этапе определяются координата и значение площади минимального сужения в тракте. Эти данные совместно с текущими значениями управляющих параметров площади $S_g(n)$ используются для вычисления объёмных скоростей и давления в различных участках речевого тракта, в т.ч. объёмной скорости $U_g(n)$, протекающей через голосовую щель, методом из [Hanson, Stevens, 2002].

На следующем этапе вычисляется объёмная скорость $U_L(n)$, протекающая через губы. Обычно для решения этой задачи используется один из двух методов. Первый метод заключается в вычислении по $T(j\omega)$ импульсной характеристики речевого тракта путём обратного преобразования Фурье с дальнейшей свёрткой этой характеристики с функцией $U_g(n)$ [Sondhi, Schroeter, 1987]. Во втором методе по $T(j\omega)$ сначала оцениваются её полюсы и вычеты для каждого полюса. Объёмная скорость у губ вычисляется средствами параллельного артикуляторно-формантного синтезатора [Lin, 1995]. На наш взгляд, эти методы не исчерпывают всех возможностей вычисления $U_L(n)$. Схема, описываемая в разделе 4, является ещё одним методом решения данной задачи.

Наконец, в последнем блоке вычисляется речевой сигнал $s(n)$ путём соответствующей фильтрации $U_L(n)$.

Как указано в разделе 1, описанная схема не позволяет учитывать влияние подсвязочной области на акустические характеристики речевого тракта. Соответствующая модификация схемы будет построена в разделе 4.

3. Взаимодействие подсвязочной области и речевого тракта

Рассмотрим математическую модель, описывающую взаимодействие подсвязочной области с передаточной функцией речевого тракта. Пусть T — передаточная функция речевого тракта, вычисленная в предположении бесконечного акустического импеданса голосовой щели, Z_{in} — входной акустический импеданс в речевой тракт со стороны голосовой щели, Z_{sub} — входной акустический импеданс в трахею со стороны голосовой щели, Z_g — акустический импеданс голосовой щели. Тогда передаточная функция T_{tr} речевого тракта с учётом конечного импеданса голосовой щели и наличия подсвязочной области определяется так [Chi, Sonderegger, 2007]:

$$T_{tr} = T \frac{Z_g}{Z_{in} + Z_g + Z_{sub}}. \qquad (1)$$

Функции T и Z_{in} определялись с помощью обобщённой схемы длинной линии по площадям поперечных сечений, измеренным с помощью магнитно-резонансной томографии речевого тракта реального диктора для английских гласных /*A, E, I, U*/ [Baer et al., 1991]. Во всех расчётах учитывалось наличие потерь на вязкое трение и теплопроводность, а также податливость стенок речевого тракта. Числовые значения всех необходимых параметров указаны в [Макаров, 2009].

В качестве модели подсвязочной области использовалась полость объёма V_l, аппроксимирующая лёгкие, которая сочленялась с однородной трубой с потерями и податливыми стенками (трахея). Импеданс полости определялся так:

$$Z_l = R_l + \rho_0 c_0^2/(j\omega V_l). \qquad (2)$$

Здесь R_l — активное сопротивление воздушному потоку в лёгких, ρ_0 — плотность воздуха, $j=\sqrt{-1}$, ω — круговая частота (рад/с). Зная Z_l и характеристики трахеи (её длину l_{tr} и площадь поперечного сечения S_{tr}, а также параметры импеданса её стенок), можно определить входной акустический импеданс Z_{sub} подсвязочных областей.

Акустический импеданс Z_g голосовой щели вычислялся по формуле, приведённой в [Сорокин, 1985]:

$$Z_g = \frac{12\mu h_g}{l_g d_g^3} + \frac{\rho_0 U_g}{S_g} + \frac{\rho_0 h_g}{S_g}. \qquad (3)$$

Здесь μ — коэффициент вязкости воздуха, h_g — толщина голосовых складок, l_g — длина голосовых складок, d_g — ширина голосовой щели, U_g — среднее за период значение объёмной скорости, протекающей через голосовую щель, S_g — среднее за период значение площади голосовой щели. Предполагалось, что голосовая щель с достаточной степенью точности аппроксимируется прямоугольником, площадь которого может быть вычислена как $S_g = d_g l_g$.

Числовые значения параметров лёгких, трахеи и голосовой щели указаны в таблице 1.

Таблица 1

Параметры лёгких, трахеи и голосовых складок

Параметр	Числовое значение
R_l	40 акуст. Ом
V_l	3000 см^3
l_{tr}	14 см
S_{tr}	3 см^2
h_g	0.5 см
l_g	1.5 см
$_{Ug}$	200 см^3/с

Тестирование описанной модели взаимодействия подсвязочной области и речевого тракта проводилось для пяти значений площади голосовой щели S_g: 0 см^2 (случай бесконечного импеданса голосовой щели), 0.04 см^2, 0.08 см^2, 0.12 см^2, 0.2 см^2 (случай максимально открытой голосовой щели). На рис. 2 представлены амплитудно-частотные характеристики четырёх гласных звуков для указанных значений площади голосовой щели. Видно, что по мере раскрытия голосовой щели появляются дополнительные резонансы и антирезонансы, а основные резонансы сдвигаются по частоте.

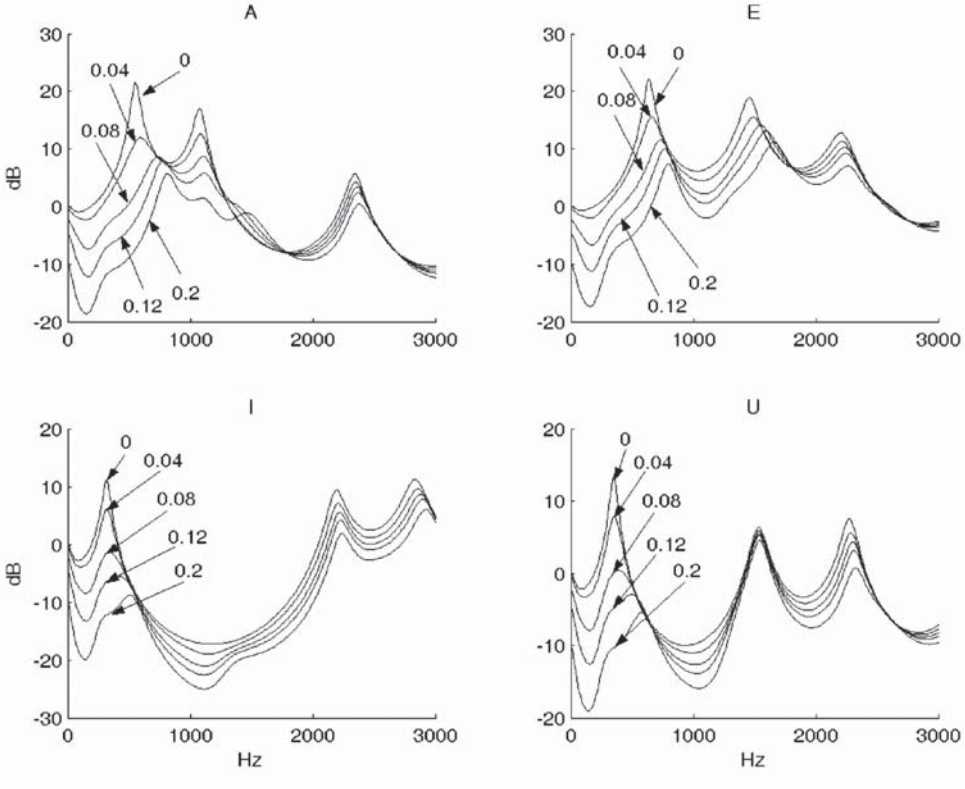


Рис. 2. Огибающие амплитудно-частотных характеристик для четырёх гласных звуков и пяти значений площади голосовой щели

Частоты основных резонансов определялись по пикам соответствующих амплитудно-частотных характеристик. Девиации F_1, F_2, F_3 (в %) по отношению к частотам $F_1^{(0)}$, $F_2^{(0)}$, $F_3^{(0)}$ для сомкнутых голосовых складок указаны в таблицах 2, 3 и 4 соответственно.

Таблица 2

Девиация первой формантной частоты

Гласный	Значение $F_1^{(0)}$ для S_g = 0 см², Гц	Девиация F_1 в % для разных значений S_g относительно $F_1^{(0)}$			
		S_g = 0.04 см²	S_g = 0.08 см²	S_g = 0.12 см²	S_g = 0.2 см²
/A/	560	8.5	30.7	39.3	44.4
/E/	634	4.5	15.0	19.6	25.6
/I/	310	0.1	6.2	40.0	61.5
/U/	348	2.7	10.9	41.1	65.8

Таблица 3

Девиация второй формантной частоты

Гласный	Значение $F_2^{(0)}$ для S_g = 0 см², Гц	Девиация F_2 в % для разных значений S_g относительно $F_2^{(0)}$			
		S_g = 0.04 см²	S_g = 0.08 см²	S_g = 0.12 см²	S_g = 0.2 см²
/A/	1074	1.8	3.6	4.4	4.4
/E/	1465	1.3	5.9	9.8	13.0
/I/	2190	0.9	1.7	1.7	2.2
/U/	1532	0.6	0.6	0.6	0.6

Таблица 4

Девиация третьей формантной частоты

Гласный	Значение $F_3^{(0)}$ для S_g = 0 см², Гц	Девиация F_3 в % для разных значений S_g относительно $F_3^{(0)}$			
		S_g = 0.04 см²	S_g = 0.08 см²	S_g = 0.12 см²	S_g = 0.2 см²
/A/	2343	0.4	0.8	1.2	1.6
/E/	2209	0.9	1.7	2.2	2.6
/I/	2845	0.8	1.6	2.0	2.8
/U/	2276	0.8	1.3	1.7	2.1

При больших значениях площади голосовой щели речевой тракт и подсвязочная область представляют собой единую акустическую систему. Поэтому в некоторых случаях было невозможно решить, какой пик соответствует резонансу речевого тракта, а какой — подсвязочной области (рис. 2, гласные /I, U/, S_g = 0.2 см², диапазон ниже 1 кГц). В этих случаях за основной резонанс формально принимался пик, лежащий на более высокой частоте, что приводило к значительным (> 60 %) девиациям по F_1 для этих гласных. Возможно, что в таких случаях вообще нельзя говорить о девиациях основных формант.

4. Модель подсвязочной области в частотно-временном артикуляторном синтезаторе

Из соотношений (1)–(3) следует, что наиболее естественно влияние подсвязочной области на акустические характеристики речевого тракта моделируется в частотной области. Это соображение подсказывает следующую схему синтеза: сначала вычисляется последовательность кратковременных спектров Фурье от функции $U_g(n)$ в окне анализа $W(n)$ со сдвигом окна в R отсчётов. Затем передаточная функция речевого тракта пересчитывается согласно соотношению (1) и умножается на кратковременный спектр Фурье объёмной скорости $U_g(n)$. Синтез сигнала во временной области осуществляется методом наложения с суммированием. На последнем этапе сигнал пропускается через фильтр, моделирующий эффект излучения на губах.

Рассмотрим описанную схему более подробно. На вход алгоритма поступают: последовательность отсчётов объёмной скорости $U_g(n)$ и набор передаточных функций $T_k(j\omega)$ и акустических импедансов $Z_{in,k}(j\omega)$, вычисляемых через равные интервалы времени Δ по площади поперечного сечения речевого тракта.

Алгоритм выглядит следующим образом.

Задаём кратковременное окно анализа $W(n)$ и сдвиг окна, равный R отсчётам. Параметр R определяет, насколько часто будет вычисляться спектр Фурье от объёмной скорости, протекающей через голосовую щель.

Вычисляем последовательность дискретных кратковременных спектров Фурье от функции $U_g(n)$:

$$\tilde{U}_r(j\omega_p) = \sum_{n=0}^{N-1} U_g(n) W(rR - n) exp(-j\omega_p n). \qquad (4)$$

В этом соотношении N — длина окна анализа в отсчётах; ω_p — дискретная сетка отсчётов по частоте ($\omega_p = 2\pi p/N$, $0 \le p \le N-1$); $\{rR\}$ — дискретная временная сетка, на которой вычисляются функции $\tilde{U}_r(j\omega_p)$, $r = 0, 1, 2, \ldots$

В каждом окне анализа вычисляем средние значения объёмной скорости через голосовую щель и площади голосовой щели. Соответствующие массивы обозначим как $u_g(r)$ и $s_g(r)$.

По $u_g(r)$, $s_g(r)$ вычисляем акустический импеданс голосовой щели $Z_{g,r}(j\omega_p)$ согласно (3). Отметим, что импеданс $Z_{g,r}(j\omega_p)$ может быть вычислен (для заданных параметров лёгких и трахеи) только один раз и в дальнейшем загружаться из файла, а не вычисляться каждый раз заново.

Интерполируем функции $T_k(j\omega)$ и $Z_{in,k}(j\omega)$, изначально заданные на временной сетке $\{k\Delta\}$, на новую временную сетку $\{rR\}$, $k, r = 0, 1, 2, \ldots$

По функциям $T_r(j\omega)$ и $Z_{in,r}(j\omega)$, $Z_{g,r}(j\omega_p)$ и $Z_{g,r}(j\omega_p)$ вычисляем передаточную функцию $\tilde{T}_r(j\omega_p)$, учитывающую влияние подсвязочной области, по формуле (1).

Вычисляем речевой сигнал методом наложения с суммированием:

$$s(n) = \sum_{r=0}^{+\infty} \left[\frac{1}{N} \sum_{p=0}^{N-1} I(j\omega_p) \widetilde{T}_r(j\omega_p) \widetilde{U}_r(j\omega_p) \exp(j\omega_p n) \right]. \quad (5)$$

В данном соотношении множитель *I* моделирует импеданс излучения на губах.

Выбор окна анализа $W(n)$ определяется двумя требованиями. С одной стороны, длительность L_t окна должна быть как можно короче для достижения хорошего разрешения по времени. С другой стороны, ширина L_w главного лепестка спектра Фурье окна $W(n)$ также должна быть как можно у́же для достижения хорошего разрешения по частоте. Хорошо известно, что эти требования противоречат друг другу в силу принципа неопределённости Гейзенберга:

$$L_t L_w \geq ½. \quad (6)$$

Точную нижнюю грань неравенству (6) доставляет временное окно Гаусса [Штарк, 2007]. Оно определяется следующей формулой:

$$W(n+1) = \exp\left(-\frac{1}{2} \left[2.5 \frac{n - N/2}{N/2} \right]^2 \right). \quad (7)$$

Иными словами, окно Гаусса обеспечивает наилучший компромисс между разрешением по времени и частоте. В силу этого обстоятельства, оно и было выбрано в качестве временного окна анализа. Длительность окна полагалась равной 10 мс (160 отсчётов при частоте дискретизации 16 кГц). Для дополнительного повышения разрешения по частоте сигнал $U_g(n)$ в окне анализа дополнялся нулями до 512 отсчётов.

Выбор параметра R определяется шириной главного лепестка L_w спектра Фурье от функции $W(n)$. Пусть $L_w = 2\pi f_c$, где f_c — значение граничной частоты главного лепестка в Гц. Тогда параметр *R* должен удовлетворять следующему неравенству [Рабинер, Шафер, 1981]:

$$R \leq 1 / (2f_c). \quad (8)$$

Несоблюдение этого условия может привести к существенным искажениям сигнала во временной области при вычислении обратного преобразования Фурье по формуле (5). В проведённом исследовании параметр *R* полагался равным 2 мс (32 отсчёта при частоте дискретизации 16 кГц).

Формула (5) является классическим соотношением для синтеза сигнала по последовательности его кратковременных спектров Фурье методом наложения с суммированием [Рабинер, Шафер, 1981]. Известны различные модификации этой формулы [Griffin, Lim, 1984; Veldhuis, He, 1996], основанные на введении кратковременных окон синтеза и различных нормировок сигнала. Мы реализовали все модификации, однако не обнаружили никакой перцептивной разницы между сигналами, синтезированными по разным схемам. Поэтому в дальнейшем мы обсуждаем только результаты, полученные на основе формулы (5).

5. Эксперименты

Данными для экспериментов послужили площади поперечного сечения, основанные на результатах ядерно-магнитной томографии речевого тракта реального диктора [Story, 2005] и включавшие в себя произношения отдельных гласных и различных звукосочетаний. В дальнейшем анализе использовались последовательности площадей сечения для гласных /A, U, I/, а также для звукосочетаний /AU/ и /IO/. Исходные акустические файлы для данных звуков были недоступны, так что мы не могли сопоставить синтезированные сигналы с их исходным звучанием.

Сначала была проверена гипотеза о перцептивной эквивалентности стандартных схем частотно-временного синтезатора и схемы из раздела 4 без учёта влияния подсвязочной области. Для этого сигналы были синтезированы с помощью схем из [Sondhi, Schroeter, 1987; Lin, 1995], а также схемы, построенной выше. Чтобы исключить влияние подсвязочной области, вместо функции $\widetilde{T}_r(j\omega_p)$ в формуле (5) использовалась функция $T_r(j\omega)$. На слух сигналы оказались неотличимыми друг от друга.

Затем были синтезированы речевые сигналы с учётом подсвязочной области, параметры для которой заимствовались из таблицы 1. Прослушивание результирующих сигналов на бюджетной бытовой аудиоаппаратуре не выявило между ними существенной разницы. При прослушивании этих же звуков на высококачественной студийной аппаратуре (мониторах KRK VXT4 или головных телефонах AKG K240 mk2, подключённых к аудиоинтерфейсу E-MU 0404 USB, и другой аппаратуре более высокого класса) разница между ними оказалась весьма заметной. При этом сигналы, синтезированные с учётом влияния трахеи, бронхов и лёгких, имели более сочный и бархатистый оттенок по сравнению с сигналами, синтезированными без подсвязочной области.

Напомним, что в данной работе не ставилась задача выяснения роли подсвязочной области для восприятия речи. Решение такой задачи требует особого исследования.

6. Обсуждение

Цель проведённого исследования заключалась в создании схемы артикуляторного синтезатора, моделирующей взаимодействие речевого тракта и подсвязочной области. Построенная схема обладает двумя существенными преимуществами перед стандартными алгоритмами артикуляторного синтеза. Первое заключается в том, что влияние трахеи, бронхов и лёгких на передаточную функцию речевого тракта моделируется в частотной области. Это позволяет учесть нелинейные функции частоты в переносном импедансе трахеи наиболее адекватным образом, без упрощающих предположений, характерных для временных синтезаторов. Вторым преимуществом является использование кратковременного анализа Фурье и метода наложения с суммированием для синтеза сигнала. Выбирая достаточно высокую частоту вычисления спектра Фурье (или, что эквивалентно, увеличивая значение параметра R), можно синтезировать быстрые изменения спектральных характеристик речевого сигнала (что практически нереализуемо в стандартных методах частотно-временного артикуляторного синтеза).

В настоящем исследовании соотношение (5) реализовано посредством вычисления обратного быстрого преобразования Фурье от спектров звуков. Возможны иные реализации этого соотношения. Например, раскладывая комплексные экспоненты в (5) по синусам и косинусам, можно осуществить синтез средствами гармонического вокодера

[McAulay, Quatieri, 1986; George, Smith, 1997]. Соотношение (5) может быть реализовано методами, разработанными в рамках модуляционной модели речевого сигнала [Potamianos, Maragos, 1999]. Преимущество подобных схем перед схемой, описанной в разделе 4, очевидно в методе синтеза речевой волны синхронно с основным тоном. В этом случае длина окна анализа *N* становится равной длительности текущего периода основного тона в отсчётах, что приводит к значительным ускорениям в вычислениях и уменьшению объёмов памяти. Например, для частоты основного тона 100 Гц и частоты дискретизации 16 кГц требуется вычислять передаточную функцию тракта и акустические импедансы лишь для 80 гармоник (до частоты Найквиста), в то время как метод, построенный в разделе 4, требует вычисления акустических характеристик тракта для 512 спектральных отсчётов.

Несмотря на то, что схема была построена для учёта взаимосвязи речевого тракта и подсвязочных областей, она может эффективно использоваться для синтеза всех явлений акустики речеобразования, наиболее адекватно моделируемых именно в частотной области. В качестве примера можно упомянуть турбулентный шум, служащий основным источником звука для фрикативных согласных. Известно [Zhang et al., 2002], что многие характеристики турбулентного шума существенно зависят от трёхмерной геометрии шумящего отверстия, наличия/отсутствия препятствия на пути воздушного потока, а также от угла падения потока на рассеивающее препятствие. Практика показывает, что адекватно моделировать эти явления значительно проще в частотной, а не во временной области [Narayanan, Alwan, 2000]. В качестве другого примера можно указать «обратное» влияние входного импеданса в речевой тракт со стороны голосовой щели на воздушный поток, протекающий через гортань. Во временной области это влияние моделируется свёрткой данного импеданса с отсчётами объёмной скорости $U_g(n)$. Технические трудности данного подхода подробно обсуждаются в [Sondhi, Schroeter, 1987]. С другой стороны, в частотной области это влияние сводится к произведению входного акустического импеданса и спектра Фурье от объёмной скорости через голосовую щель.

Литература

1. Сорокин В.Н. Теория речеобразования. М.: Радио и Связь, 1985. 312 с.

2. Stevens K. Acoustic Phonetics. The MIT Press, 2000. 614 p.

3. Chi X., Sonderegger M. Subglottal Coupling and its Influence on Vowel Formants // J. Acoust. Soc. Amer. 2007. Vol. 122, P. 1735–1745.

4. Сорокин В.Н. Синтез речи. М.: Наука, 1992. 392 с.

5. Assmann P., Katz W. Synthesis Fidelity and Time-Varying Spectral Change in Vowels // J. Acoust. Soc. Amer. 2005. Vol.117. P. 886–895.

6. Морозов В.П. Искусство резонансного пения. М., 2002. 496 с.

7. Kawahara H., Masuda-Katsue I. and Cheveigne A.de. Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous — Frequency Based F0 Extraction: Possible Role of a epetitive Structure in Sounds // Speech Communication. 1999. Vol.27. P. 187–207.

8. Баден П., Макаров И.С., Сорокин В.Н. Алгоритм вычисления площадей поперечного сечения речевого тракта // Акуст. журнал. 2005. Т. 51. No. 1. С. 52–58.

9. Story B., Titze I. andHoffman E. Vocal Tract Area Functions for an Adult Female Speaker Based on Volumetric Imaging // J. Acoust. Soc. Amer. 1998. Vol.104. P. 471–487.

10. Сорокин В.Н., Макаров И.С. Определение пола диктора по голосу // Акуст. журнал. 2008. Т. 54. No. 4. С. 659–668.

11. Макаров И.С. Аппроксимация речевого тракта коническими рупорами // Акуст. журнал. 2009. Т. 55. No. 2. С. 256–265.

12. Hanson H., Stevens K. A Quaisarticulatory Approach to Controlling Acoustic Source Parameters in a Klatt-Type Formant Synthesizer Using HLsyn // J. Acoust. Soc. Amer. 2002. Vol. 112. P. 1158–1182.

13. Sondhi M.M., Schroeter J. A Hybrid Time-Frequency Domain Articulatory Synthesizer // IEEE Trans. Acoust., Speech, Signal Process. 1987. Vol. ASSP-35. No. 7. P. 955–967.

14. Lin Q. A Fast Algorithm for Computing the Vocal Tract Impulse Response from the Transfer Function // IEEE Trans. Speech, Audio Process. 1995. Vol. 3. No. 6. P. 449–457.

15. Штарк Г. Применение вейвлетов для цифровой обработки сигналов. М.: Техносфера, 2007. 192 с.

16. Рабинер Л., Шафер Р. Цифровая обработка речевых сигналов. М.: Радио и Связь, 1981. 496 с.

17. Griffin D., Lim J. Signal Estimation From Modified Short-Time Fourier Transform // IEEE Trans. Acoust., Speech, Signal Process. 1984. Vol. 32. No. 2. P. 236–243.

18. Veldhuis R., He H. Time-Scale Pitch Modification of Speech Signals and Resynthesis From the Discrete Short-Time Fourier Transform // Speech Communication. 1996. Vol. 18. P. 257–279.

19. Story B. A Parametric Model of the Vocal Tract Area Function for Vowel and Consonant Simulation // J. Acoust. Soc. Amer. 2005. Vol. 117. P. 3231–3254.

20. McAulay K., Quatieri T. Speech Analysis / Synthesis Based On a Sinusoidal Representation of Speech // IEEE Trans. Acoust., Speech, Signal Process. 1986. Vol. 34. P. 744–754.

21. George E., Smith M. Speech Analysis / Synthesis and Modification Using an Analysis — by — Synthesis / Overlap — Add Sinusoidal Model // IEEE Trans. Speech, Audio Process. 1997. Vol.5. No.6. P. 389–406.

22. Potamianos A., Maragos P. Speech Analysis and Synthesis Using an AM-FM Modulation Model // Speech Communication. 1999. Vol.28. P. 195–209.

23. Narayanan S., Alwan A. Noise Source Models for Fricative Consonants // IEEE Trans. Speech, Audio Process. 2000. Vol.8. No.2. P. 328–344.

24. Zhang Zh., Mongeau L. andFrankel S. Broadband Sound Generation by Confined Turbulent Jets // J. Acoust. Soc. Amer. 2002. Vol.112. P. 677–689.

Горбунов К.С.,

Макаров И.С.,
кандидат технических наук, старший научный сотрудник
Института Проблем Передачи Информации им. А.А. Харкевича РАН.
speechprod_mak@mail.ru

Базы данных речи русских детей «INFANTRU» и «CHILDRU»

Ляксо Е.Е., Фролова О.В., Громова А.Д., Гайкова Ю.С., Куражова А.В., Романова О.Д., Богорад М.А., Остроухов А.В., Соловьёв А.Н., Смит Н.Ю

Введение

Базы данных звуков и речи русских детей в возрасте от 3 месяцев до 7 лет «INFANTRU» и «CHILDRU» являются первыми базами детской речи на материале русского языка. База звуков и речи детей первых трёх лет жизни «INFANTRU» содержит лонгитюдные записи речи/звуков 187 детей, представленные пятиминутными фрагментами и отдельными сигналами, произнесёнными ребёнком в различных психоэмоциональных состояниях. Записи выполнены в условиях спонтанного произнесения и взаимодействия ребёнка с матерью. База «CHILDRU» содержит записи речи 150 детей в возрасте от 4 до 7 лет и является продолжением базы «INFANTRU». Запись речевого материала детей 4–7 лет выполнена в ситуациях взаимодействия со взрослым: спонтанная речь, ответы на вопросы, чтение, стихи или пересказ рассказа, счёт и алфавит, игра. В базу «CHILDRU» внесена информация о детях, о матери, условиях записи и записывающей аппаратуре. Диалоги детей с взрослыми, отдельно произнесённые фразы детей, слова, слоги и звуки интерпретированы в орфографическом представлении. Слова, которые дети произносили с ошибками, транскрибированы в символах МФА. Для работы с базами данных созданы программы, позволяющие выбрать речевой материал в зависимости от возраста и ситуации для всех и для каждого из детей.

Материал, объём и структура современных речевых баз данных

В зависимости от целей и задач исследований речевые базы данных, как правило, характеризуются следующими параметрами:

- объём базы (количество дикторов/респондентов);
- речевой материал: читаемая речь — предопределённые слова-команды (фонетически богатые предложения), вызванная речь, спонтанная речь;
- тип канала связи: стационарная телефонная сеть, мобильная связь, широкополосный канал и т.д.

Современные речевые базы данных создаются, главным образом, для решения задач автоматического распознавания речи (например, проекты: SpeechDat-I, SpeechDat-II, SpeechDat-E, SpeechDat-Car, Speecon) и вери-

фикации говорящего [Викторов и др., 1999; Galunov et.al., 2002]. Они включают фонограммы взрослых дикторов. Существующие общепринятые технологии распознавания речи (Hidden Markov Models) определяют формат таких баз данных: большое количество различных дикторов (несколько тысяч), статистически достаточное представление всех фонем (аллофонов) данного языка, представительность дифонов и трифонов. Просодическая организация высказывания, построение диалоговой речи, как правило, не рассматриваются.

Базы детской речи

На сегодняшний день существует ограниченное количество баз данных детской речи, или они имеют узкоспециализированную направленность, что обусловлено недостаточной изученностью акустики детского речевого сигнала и, как следствие, проблематичностью применения существующих технологий для распознавания детской речи. Тем не менее, преимущественно на материале английского языка создаются базы данных, позволяющие разрабатывать подходы к автоматическому распознаванию детской речи [Potamianos, Shrikanth, Sungbok,1997].

Автоматическое распознавание речи детей школьного возраста имеет большое значение для создания компьютерных программ, позволяющих моментально и объективно оценивать и корректировать процесс чтения. В связи с этим формируются речевые базы данных, содержащие читаемый детьми материал. Такие базы созданы для детей 8–15 лет, воспитывающихся в англоязычной среде [Hagen, Pellom, Cole, 2003], итальянских детей 7–13 лет [Cosi, Pellom, 2005; Giuliani, Gerosa, 2005], бельгийских детей — носителей голландского языка [Cleuren, Duchateau, Ghesqui`ere, Van Hamme, 2008]. Эти базы наряду с нормативными данными содержат речевой материал детей, испытывающих трудности в процессе освоения навыка чтения [Cleuren, Duchateau, Ghesqui`ere, van Hamme, 2008].

Для решения вопросов о возможности речевого взаимодействия ребёнка с компьютером, для задач диагностики и исправления речевых нарушений большое значение имеют базы данных, содержащие вызванную, спонтанную и эмоциональную речь детей различного возраста. Такие базы данных созданы на материале различных языков [Vicsi et al, 1999; Batliner et al, 2005; Shobaki, Hosom, Cole, 2000]. Существует речевая база нормативно развивающихся и имеющих речевые нарушения венгерских детей 5–10 лет [Csatari, Bakcsi, Vicsi, 2006], база, содержащая спонтанные диалоги шведских детей 8–15 лет [Bell et al, 2005]. В детской речевой базе PF STAR собрана читаемая речь и имитационные высказывания английских, немецких и шведских детей 4–12 лет, чтение и повторение высказываний на английском языке 10–11-летними немцами, итальянцами и шведами, а также спонтанная и эмоциональная речь английских и немецких детей 4–14 лет [Batliner et al, 2005].

С целью создания мультимедийной обучающей системы для детей, имеющих речевые и слуховые нарушения, сформирована специализированная база речи 5–10-летних английских, шведских, словенских и венгерских детей [Vicsi et al, 2000]. Она содержит материал, представляющий собой читаемые или повторяемые детьми одно-, двух-, трёхсложные слова и фразы с определёнными гласными и согласными звуками.

База вокализаций и речи англоязычных детей первых четырёх лет жизни «LENA» была собрана в результате комплексных исследований роли окружения в речевом развитии

ребёнка. «LENA» содержит более 45 часов аудиозаписей речевого взаимодействия в 329 семьях с различным социально-экономическим статусом (Gilkerson, Coulter, Richards, 2008).

Широко известна база данных «CHILDES» [MacWhinney, Snow, 1985; MacWhinney, 1995], используемая преимущественно лингвистами – исследователями детской речи. Она содержит расшифрованные материалы, разделённые на отдельные корпусы. Так, «CHILDES» включает «корпус английского языка», «германский корпус» (голландский, датский, немецкий и шведский языки), «корпус романских языков» (каталанский, французский, итальянский, португальский, испанский языки), а также корпус «другие языки» (кантонский диалект китайского языка, мандаринский диалект китайского языка, эстонский, греческий, венгерский и др.). Отдельно представлены данные об освоении языка детьми-билингвами, а также клинические данные. При помощи системы «CHILDES» расшифровывается и речь русскоязычных детей [Доброва, 2009].

База данных звуков и речи русскоязычных детей 0–3 лет жизни «INFANTRU»

Коллективом сотрудников группы по изучению детской речи Биолого-почвенного факультета СПбГУ создана база данных звуковых и речевых сигналов детей в возрасте 3–36 месяцев жизни «INFANTRU» [Лаксо и др., 2005; Lyakso et al, 2007]. В основу организации материала в базе положены следующие положения:

— достаточная выборка информантов и их звуковых сигналов;

— учёт особенностей вокализаций детей первого года жизни (плач, гуление, лепет);

— специфика психоэмоционального состояния ребёнка.

Создание базы данных «INFANTRU» осуществлено в период с 1999 по 2005 год. База «INFANTRU» содержит записи вокализаций и речевых сигналов 187 детей. 99 детей воспитывались в условиях семьи, 88 детей — в условиях дома ребёнка.

Домашние дети родились и проживали в Санкт-Петербурге, их родители родились или проживают в Санкт-Петербурге не менее 10 лет. 83 ребёнка на момент записи были здоровы по заключению неонатолога и педиатра, 16 детей имели неврологически отягощённый анамнез (в т.ч. 7 детей — тяжёлые неврологические нарушения). Для 76 детей записи проведены в лонгитюде (на протяжении первого года жизни — для 32 детей, второго года жизни — для 19 детей, третьего года жизни — для 11 детей; на протяжении трёх лет жизни — для 14 детей). Общее время записи — 70 часов.

Записи звуковых и речевых сигналов детей произведены в следующих ситуациях: ребёнок находится один, взаимодействует с матерью в сценариях: «лицом к лицу», игра, чтение, спонтанное взаимодействие. Каждый фрагмент записи, соответствующий одной ситуации в данном возрастном срезе, длится до 5 минут. Для детей первого года жизни представлены записи отдельных сигналов, произносимых в различных эмоциональных состояниях.

Для каждого звукового фрагмента указана информация о ребёнке, матери, условиях записи и используемой аппаратуре. Информация о ребёнке включает следующие сведения: имя, пол, дата и место рождения, срок гестации, протекание беременности и родов, каким по счёту является ребёнок в семье, экономическая и социальная ситуация в семье, данные о психомоторном развитии ребёнка по анкетам шкал KID и RCDI (адаптированные для детей Северо-запада опросники). Информация о матери: время проживания в Санкт-Петербурге, образование, наличие или отсутствие хронических заболеваний, возраст на момент рождения ребёнка. Информация о звукозаписи: возраст ребёнка в месяцах, ситуация записи, эмоциональное состояние ребёнка, условия записи и используемая аппаратура, имя файла.

Записи звуковых сигналов 88 детей 1,5–3 лет, воспитывающихся в условиях дома ребёнка, осуществлены в ситуации взаимодействия ребёнка с экспериментатором. Среди них 54 ребёнка первого года жизни (для 46 детей — лонгитюд), 7 детей второго года жизни (для 3 детей – лонгитюд), 27 детей третьего года жизни (для 13 детей — лонгитюд). Отличительной особенностью детей, воспитывающихся в доме ребёнка, является практически полное отсутствие звуков, поэтому запись одних и тех же детей проводилась по нескольку раз (без строгого соблюдения интервала записей в лонгитюде) и на протяжении времени, превышающего пятиминутный интервал.

Все речевые файлы представлены в формате Windows PCM, 22050 Гц, 16 бит. Для облегчения работы с базой данных создана программа VDB.EXE. Она представляет собой оболочку для поиска звуковых и речевых файлов внутри базы данных и позволяет выбирать и прослушивать речевые записи по следующим признакам: номер ребёнка, пол ребёнка, номер ребёнка в семье, заболевание ребёнка, возраст матери на момент рождения ребёнка, возраст ребёнка, полная или неполная семья, воспитатель ребёнка, ситуация записи, эмоциональное состояние ребёнка (рис.1, 2).

Насколько нам известно, аналогов подобных баз не существует. База «INFANTRU» может быть использована психологами, лингвистами, специалистами в сфере речевых технологий при выполнении научных исследований по речевому онтогенезу и создании обучающих программ.

База данных речи русскоязычных детей 4–7 лет жизни «CHILDRU»

База «CHILDRU» содержит речевой материал для 142 детей, воспитывающихся в условиях семьи, и 8 детей из дома ребёнка. Общее время записи — более 20 часов. Запись материала проведена в 2006–2008 гг. База содержит речь детей, звуковой и речевой материал которых включён в базу «INFANTRU», а также звукозаписи речи детей с 4-летнего возраста, отсутствующих в базе «INFANTRU».

Лонгитюдная запись речи детей проведена с интервалом в 6 мес. Каждая запись сопровождается подробным протоколом и/или параллельной записью поведения детей на видеокамеру. Запись речевого материала детей проведена в следующих ситуациях взаимодействия со взрослым (матерью или экспериментатором): спонтанная речь, ответы на вопросы, чтение, стихи или пересказ рассказа, счёт и алфавит, игра.

В базу внесена информация о детях, семье, условиях записи. Информация о ребёнке включает следующие сведения: имя ребёнка (сокращённое обозначение имени и фамилии), пол, дата рождения, место проживания, каким по счёту является ребё-

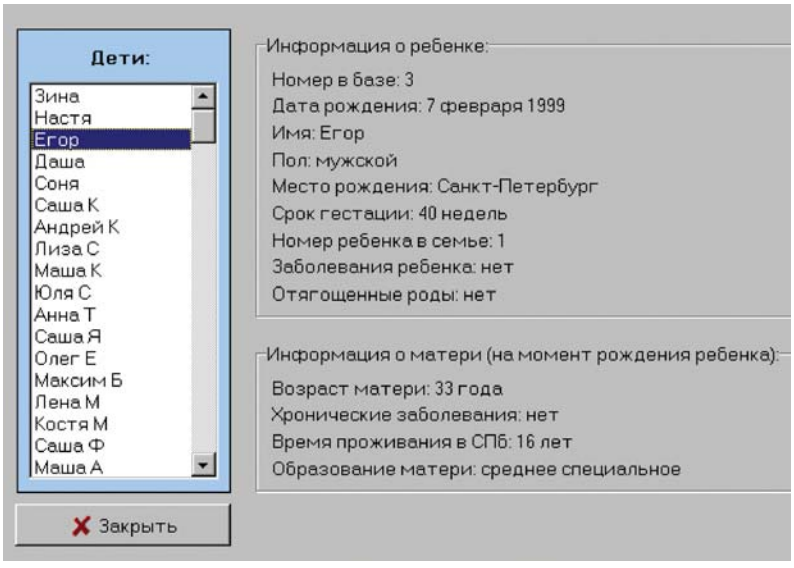


Рис. 1. Оболочка программы для поиска материала в базе данных. А — условия записи, Б — информация о ребёнке

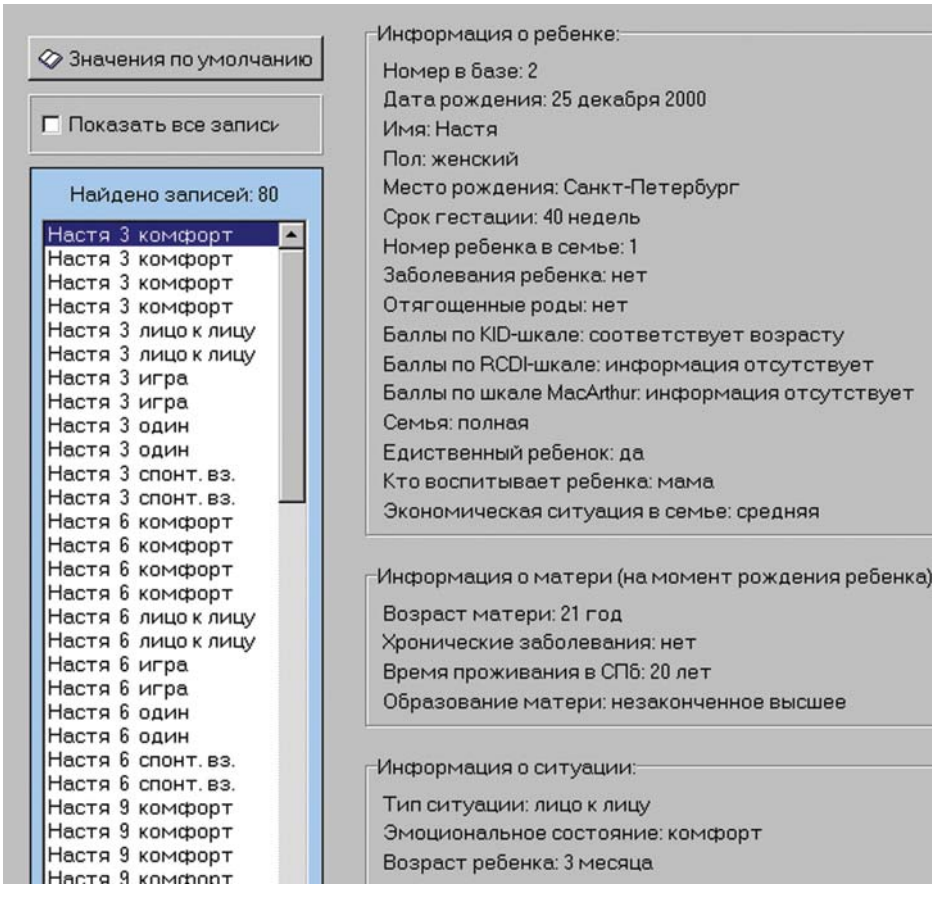


Рис. 2. Полная информация о выбранном в базе ребёнке, его матери и условиях записи

нок в семье, наличие или отсутствие братьев или сестёр, наличие или отсутствие пренатальных и хронических заболеваний, посещает/не посещает дошкольное заведение (детский сад), какой детский сад ребёнок посещает (обычный или логопедический), соответствует/не соответствует возрастным нормам развитие ребёнка, номер ребёнка в базе «INFANTRU». Информация о семье: время проживания матери ребёнка в Санкт-Петербурге на момент рождения ребёнка, образование матери, полная/неполная семья, кто воспитывает ребёнка (мама, папа, бабушка, няня). Информация об условиях записи: место записи, записывающая аппаратура, ситуация записи.

Речевой материал в базе «CHILDRU» представлен в виде оригинальных файлов длительностью до 5 мин., отражающих ситуацию: спонтанная речь, чтение, ответы на вопросы, стихи и рассказ, счёт и алфавит, игра. Оригинальный файл, наряду с речью ребёнка, может содержать речь мамы, экспериментатора и других детей, а также различные шумы в игровых ситуациях.

Из оригинального файла выбираются: фразы; диалоги ребёнка со взрослым; вопросы; отдельно произносимые звуки и слоги; слова, состоящие из одного, двух, трёх, четырёх, пяти и более слогов; слова, произнесённые детьми с ошибками (связанными с различными вариантами замен, пропусков и перестановок фонемы /r/ в словах; ошибки, обусловленные пропуском, заменой и перестановкой других фонем или слогов в слове; неправильное построение фразы); чтение слов и фраз.

Особенностью базы «CHILDRU» является то, что весь речевой материал сопровождается текстовыми файлами, в которых приведено орфографическое описание диалогов взрослый–ребёнок, отдельных детских фраз и слов. Слова, которые дети произносили с ошибками, описаны в терминах МФА (Международного фонетического алфавита). Речевые файлы представлены в формате Windows PCM, 22050 Гц, 16 бит. Структура базы «CHILDRU» повторяет собой структуру предшествующей базы «INFANTRU». Однако в связи с большим объёмом содержащегося речевого материала проведена работа по усовершенствованию базы, направленная на создание программы, максимально облегчающей пользователю работу с базой. Программа работает в двух режимах — просмотра и редактирования. В режиме редактирования пользователь может добавлять звуковой материал в базу. В режиме просмотра осуществляется поиск информации в базе по всем признакам: информация о ребёнке, матери и условиях записи.

Речевой материал, содержащийся в базе, сведения о каждом информанте и условиях записи находит широкое применение в научных исследованиях, посвящённых изучению фундаментальной проблемы — становления процесса речеобразования и формирования основ русского языка в онтогенезе.

Примеры обработки информации, содержащейся в базах данных

Одним из направлений исследований речевого онтогенеза является изучение акустического аспекта речи детей. Наличие лонгитюдных записей вокализаций и слов, реализованных детьми, позволило провести исследование формирования акустического облика гласных. В серии наших работ рассматривается становление системы признаков речевого сообщения, характерных для речи взрослого и обеспечивающих доступность для понимания взрослым значения детских высказываний.

На двухформантной плоскости значения двух первых формант гласного [a] в вокализациях и словах в исполнении детей находятся в области расположения соответствующих значений гласного [a] в речи взрослого. Значения первой форманты гласных [u], [i] превышают соответствующие значения для речи взрослого. Более того, в 24-месячном возрасте значения обеих формант гласного [u] расположены в более высокочастотной области (см. рис.3). Различия между формантными частотами гласных [a], [i], [u] в вокализациях и словах детей от 3 до 60 месяцев представлены в таблице 1 [Lyakso, Frolova, 2007].

Таблица 1

Различия между значениями формантных частот гласных [a], [i], [u], отнесённых носителями русского языка к соответствующим категориям гласных

Возраст, месяцы	a\u	a\i	i\u
3	–	–	–
12	–	–	–
24	–	–	–
36	F1: p<0.05; F2: p<0.01	F2: p<0.05	F2: p<0.001
48	F2: p<0.01	F1: p<0.05; F2: p<0.001	F2: p<0.001
60	F1: p<0.01; F2: p<0.01	F1: p<0.001; F2: p<0.001	F2: p<0.001

Материал, содержащийся в базе, позволил провести сравнительное исследование лексикона здоровых детей, воспитывающихся в семье (норма, n=27); детей, имеющих неврологические нарушения и воспитывающихся в семье (риск, n=6); детей из дома ребёнка (депривация, n=7) [Ляксо, Столярова, 2008; Ляксо, Столярова, Охарева, 2008].

Анализ частоты встречаемости в лексиконе детей слов, состоящих из различного числа слогов, выявил преобладание слов, состоящих из двух слогов, в лексиконе детей групп нормы и риска в 4 года и 4 года 6 мес., что соответствует данным, приведённым в частотном словаре русского языка. Однако распределение слов с различным количеством слогов у детей, входящих в группы нормы (0.29: 0.32: 0.27: 0.1: 0.02 — соответственно распределение слов с 1:2:3:4:5 слоговой структурой) и риска (0.32: 0.52: 0.11:0.33:0), различается. Лексикон детей группы нормы по этому показателю является более сложным.

Для 4-летних детей группы депривации характерно значимое преобладание слов, состоящих из одного слога (0.67), по сравнению со словами с большим количеством слогов (рис. 4а). Данное соотношение изменяется в сторону увеличения двуслоговых слов в 4 года 6 мес. (0.22 и 0.51 — частота встречаемости двуслоговых слов в лексиконе детей 4 лет и 4 лет 6 мес. соответственно) (рис. 4б). Дети группы нормы и риска в 4 года употребляют слова, состоящие из 4 и 5 слогов; в группе депривации слова, состоящие из 5 слогов, отсутствуют.

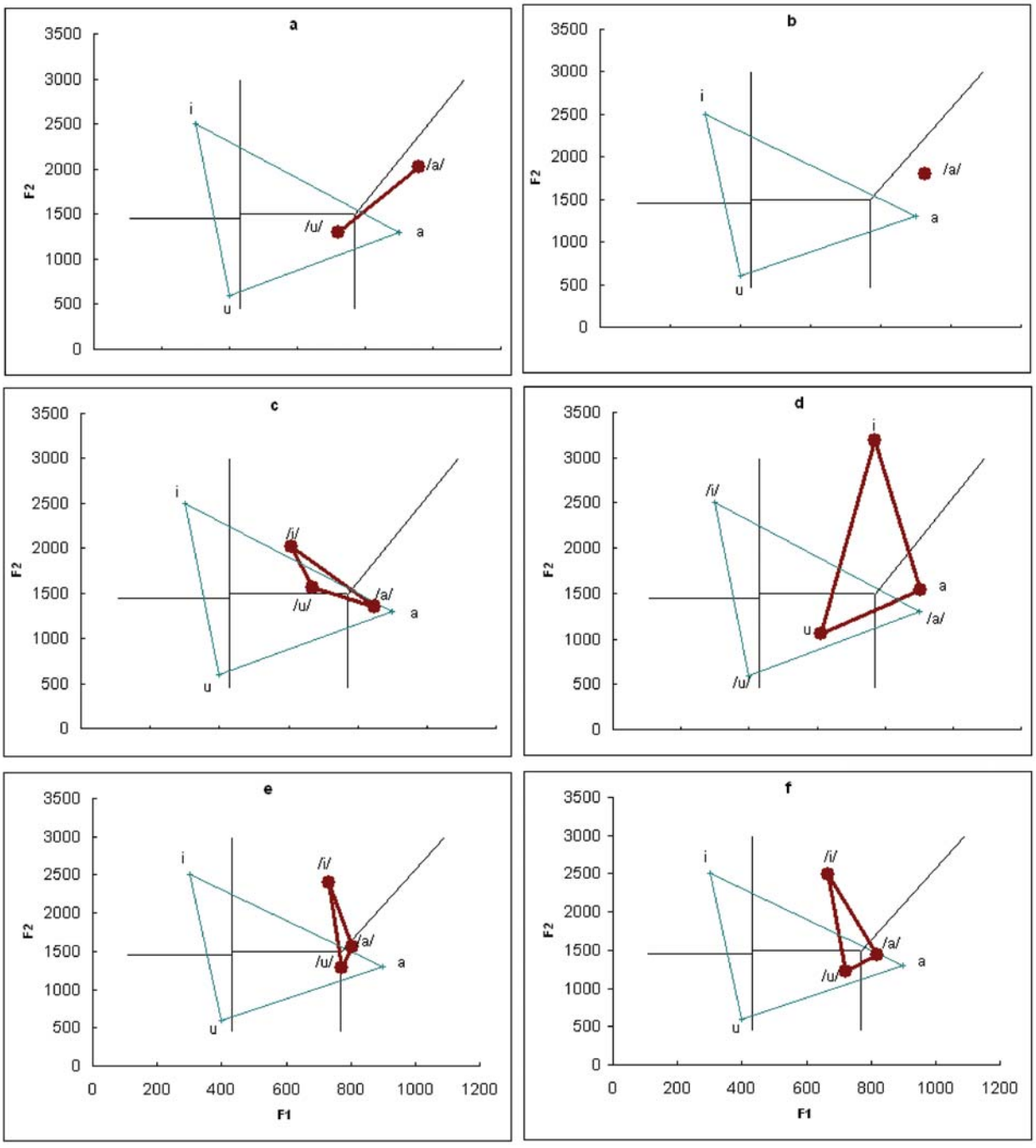


Рис. 3. Формантные треугольники гласных с вершинами [а], [u], [i] (медианные значения) для детей от 3 до 60 месяцев (тёмные линии) и для речи взрослого.
(Линиями обозначены фонемные границы восприятия гласных русского языка [Слепокурова, 1979];
по горизонтальной оси — значения первой форманты (F1) в Гц, по вертикальной оси — значения второй форманты (F2) в Гц;
a — данные для детей 3 мес., b — 12 мес., c — 24 мес.; d — 36 мес., e — 48 мес., f — 60 мес.)

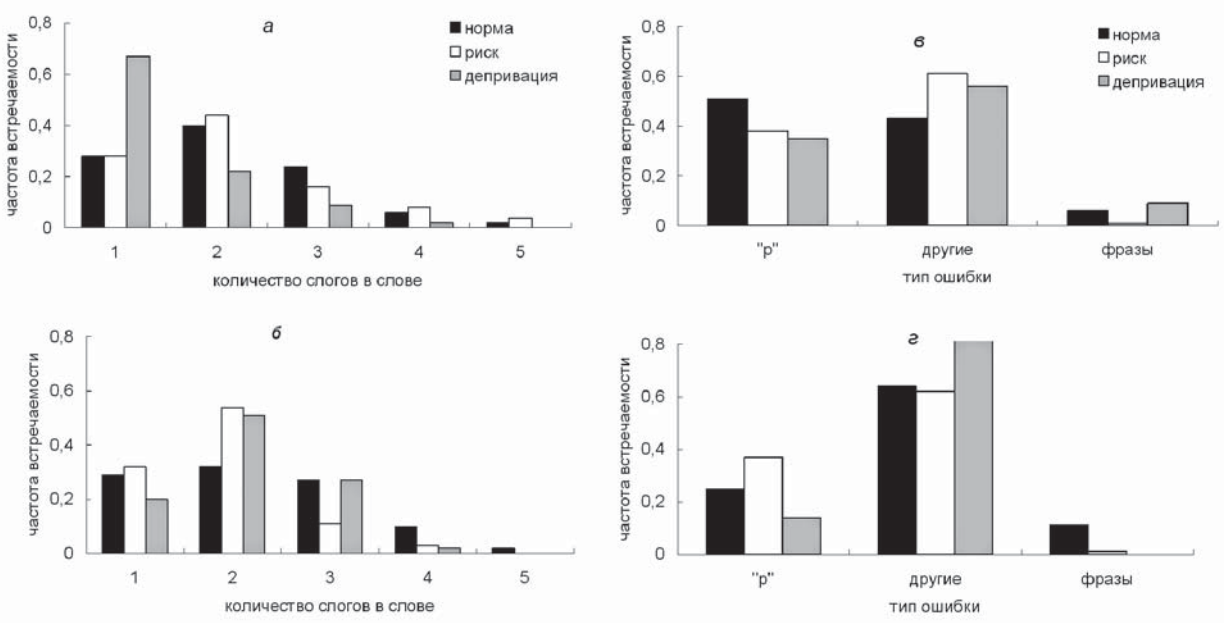


Рис. 4. Слоговой состав слов и ошибки в активном лексиконе детей 4 лет и 4 лет 6 мес. в группах нормы, риска и депривации
(а, б — частота встречаемости слов с разным количеством слогов в лексиконе детей трёх групп; в, г — артикуляционные и грамматические ошибки; а, в — данные для детей 4 лет; б, г — то же для детей 4 лет 6 мес.;
на гистограммах: чёрный цвет — данные для детей группы нормы, белый — группы риска, серый — группы депривации; «р» — слова, содержащие разные варианты ротацизма, «другие» — все остальные артикуляционные ошибки; «фразы» — грамматические ошибки, связанные с построением фраз)

По сравнению с детьми группы нормы дети группы риска употребляют больше одно- и двуслоговых слов и допускают больше ошибок при их произнесении, а у детей группы депривации эти показатели ещё более выражены.

Дети группы нормы наряду с ответами, состоящими из одного слова (0.46 и 0.26 — соответственно в 4 года и 4 года 6 мес.) и одной простой фразы (0.42 и 0.48 — соответственно в 4 года и 4 года 6 мес.), отвечают двумя и/или несколькими фразами. Они используют при общении сложноподчинённые конструкции, количество которых увеличивается к 4 годам 6 мес. (рис. 5а,б). Структура ответных реплик более сложная, чем у детей группы риска: в 4 года – за счёт употребления нескольких фраз и сложных фраз, отсутствующих у детей группы риска; в 4 года 6 мес. — за счёт сложных фраз, уменьшения количества ответов типа «да–нет» и ответов, являющихся повторением части вопроса взрослого. Дети группы депривации преимущественно отвечают односложно (0.66 и 0.42 — соответственно в 4 года и 4 года 6 мес.).

Для детей группы депривации характерен пропуск реплик (молчание в ответ на реплику взрослого), выявленный в 56% диалогов всех 4-летних детей и в 42.8% диалогов ребёнка в возрасте 4 года 6 мес. Количество пропущенных реплик составляет 48±22% — для 4-летних детей; 19.6±2,5% — для

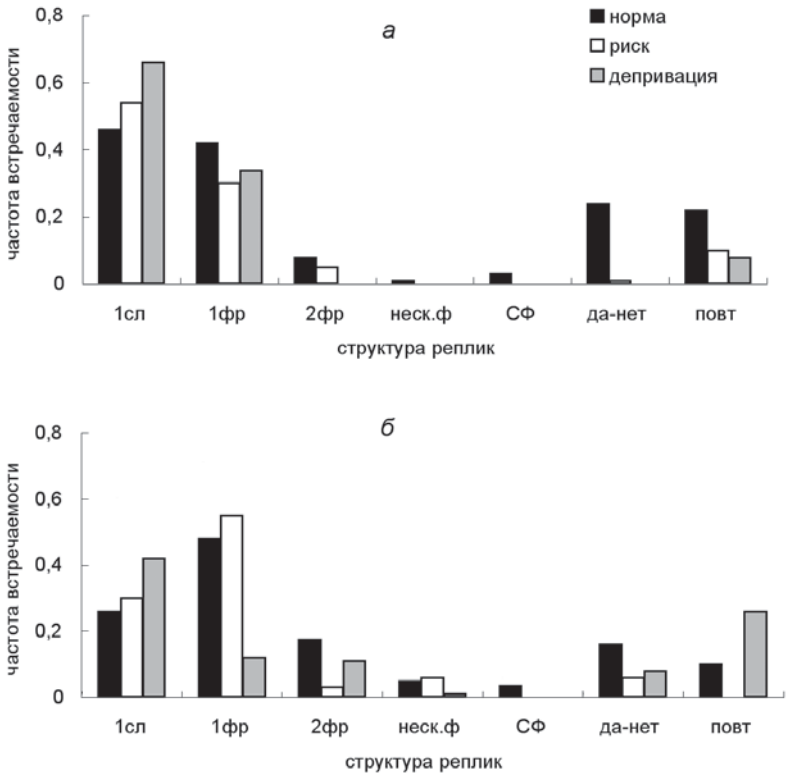


Рис. 5. Структура ответных реплик ребёнка в диалоге со взрослым (а — данные для детей в возрасте 4 лет, б — то же для детей в 4 года 6 мес.; по горизонтальной оси: 1 сл — одно слово, 1 фр — одна фраза, 2 фр — две фразы, неск.ф — несколько простых фраз, СФ — сложная фраза, да–нет — ответная реплика одним словом «да» или «нет», повт — повторение в ответной реплике части реплики взрослого)

ребёнка в 4 года 6 мес. В группе нормы пропущенные реплики не выявлены, в группе риска — у ребёнка Н. в 4 года (40% диалогов, 20% реплик) и ребёнка Т. в 4 года 6 мес. (33% диалогов, 28% реплик).

Отличия между детьми групп нормы и депривации выявлены и по показателю относительного количества артикуляционных ошибок в словах ответных реплик. У детей обеих групп количество ошибок возрастает в словах с большим количеством слогов (рис. 6), но дети группы депривации делают значимо больше ошибок при произнесении всех слов, а слова, состоящие из 4 и 5 слогов, практически постоянно произносят неправильно (0.75 и 1.0 — частота встречаемости ошибок в словах из 4 и 5 слогов).

Данное исследование показало, что в лексиконе детей групп нормы и риска преобладают слова, состоящие из двух слогов, при наличии слов с более сложной слоговой структурой, но у детей группы риска частота встречаемости слов с меньшей слоговой структурой выше, чем у детей группы нормы.

У детей группы депривации в лексиконе отсутствуют слова, состоящие из пяти слогов. Частота встречаемости слов, состоящих из одного и двух слогов, высокая; в большом

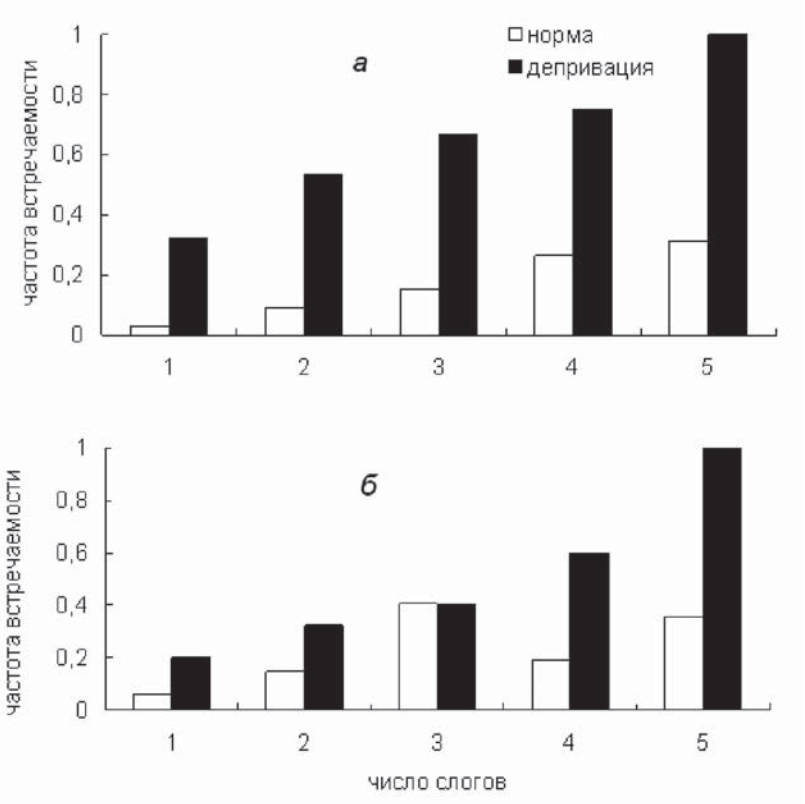


Рис. 6. Относительное число артикуляционных ошибок в словах с одинаковой слоговой структурой в ответных репликах детей групп нормы и депривации (а — данные для детей в возрасте 4 года, б — то же для детей в 4 года 6 мес.; на гистограммах: белый цвет — данные для детей группы нормы, чёрный — группы депривации)

количестве слов содержится по нескольку артикуляционных ошибок. Дети группы нормы употребляют как простые, так и сложные фразы; дети группы риска — преимущественно простые фразы, дети группы депривации — только простые фразы. Диалоги детей группы нормы характеризуются большей длительностью, чем диалоги детей групп риска и депривации, как за счёт большего количества реплик, так и за счёт усложнения ответа. Ответные реплики детей группы нормы имеют более сложную синтаксическую организацию, чем у детей групп риска и депривации. Для детей группы депривации характерна замена вербальных реплик простыми жестами и пропуск реплик.

Анализ структуры реплик детей в диалогах со взрослым собеседником и в диалогах между собой показал, что в обеих возрастных группах в диалогах с взрослым собеседником около 40% ответов детей содержит одно слово, в диалогах детей такие ответы составляют 6% в возрасте 4 года и 4% в возрасте 5 лет (рис.7). В диалогах между собой дети употребляют более распространённые реплики (состоящие из нескольких фраз). Характер распределения частоты встречаемости реплик из одной, двух и нескольких фраз одинаков как в диалогах детей со

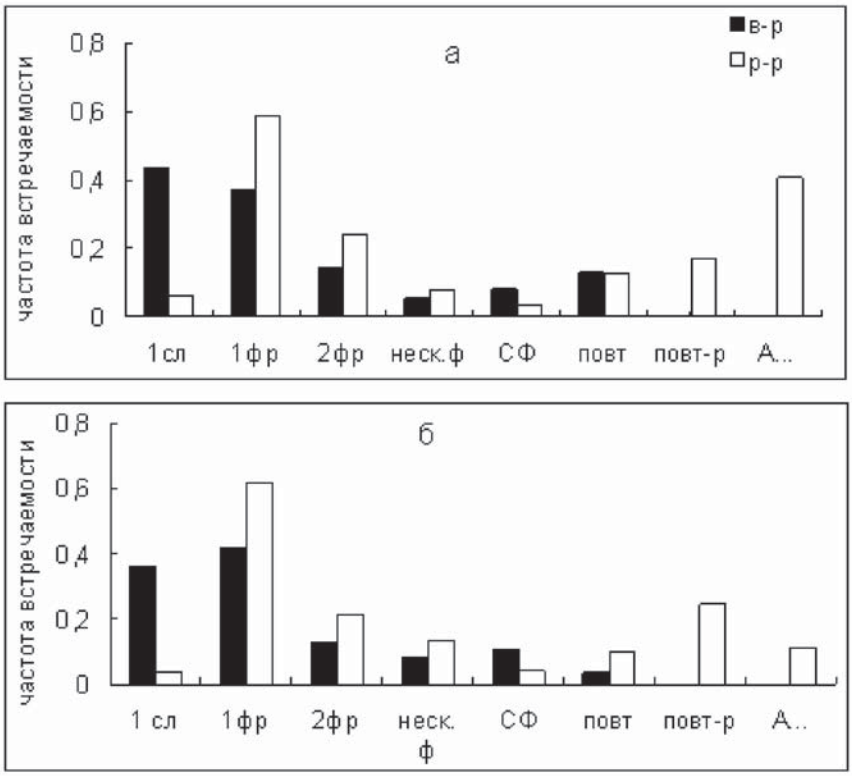


Рис. 7. Структура реплик детей в диалогах со взрослыми собеседниками (в–р) и сверстниками (р–р)
(а — данные для детей в возрасте 4 года, б — то же для детей в возрасте 5 лет;
по горизонтальной оси: 1 сл — одно слово, 1 фр — одна фраза, 2 фр — две фразы,
неск.ф — несколько фраз, СФ — сложная фраза, повт — повторение в ответной реплике части инициирующей реплики, повт-р — повтор словосочетаний внутри одной реплики,
А... — начало реплики со слов;
на гистограммах: чёрный цвет — данные для реплик в диалоге со взрослым,
белый — с ребёнком)

взрослым, так и при общении со сверстниками. Наибольшее число реплик состоит из одной фразы.

В диалогах со взрослым дети чаще употребляют распространённые и сложноподчинённые предложения. Реплики-повторы наблюдаются как в диалогах со взрослыми, так и при общении детей между собой. В реплике-повторе речи сверстника дети копируют не только словесную конструкцию, но и интонацию, зачастую усиливая её выраженность. Характерным признаком речевого общения детей между собой является наличие в одной реплике ребёнка повторов слова, нескольких слов или даже фраз. Такие реплики составляют 17% у детей в возрасте 4 года и 24% — в возрасте 5 лет. Число повторов в одной реплике может составлять от одного до трёх.

Осуществлён анализ активного лексикона по частоте встречаемости слов с разным количеством слогов, реализованных всеми детьми в возрасте 4–7 лет. Записи данного речевого материала содержатся в базе «CHILDRU» (рис.8).

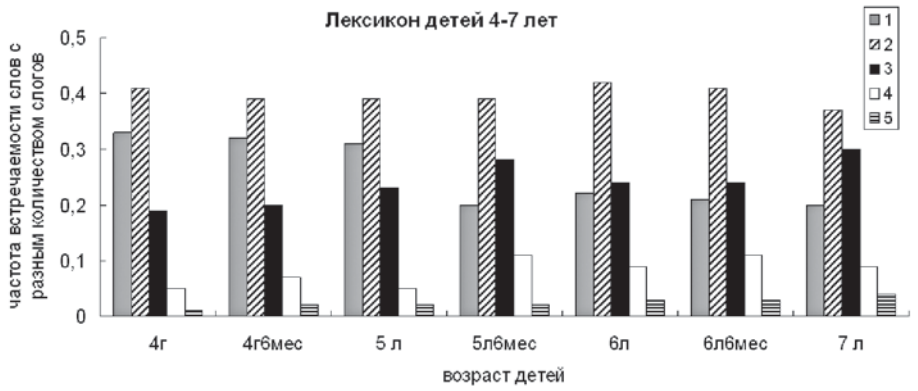


Рис. 8. Анализ лексикона детей 4–7 лет
(обозначения на гистограмме — количество слов в зависимости от количества слогов: серый — из одного слога, наклонная штриховка — из двух слогов, чёрный — из трёх, белый — из четырёх, горизонтальная штриховка — из пяти и более слогов)

Для всех детей, записи которых присутствуют в финальной версии базы данных, проанализирован лексикон. Проведён подсчёт количества слов, состоящих из 1, 2, 3, 4, 5 и более слогов, у детей в возрастных срезах 4 года, 4 года 6 мес., 5 лет, 5 лет 6 мес., 6 лет, 6 лет 6 мес., 7 лет.

- Во всех возрастных срезах в лексиконе детей преобладают слова из 2 слогов: 42% слов — в 4 года, 40% — в 5 лет, 43% — в 6 лет и 37% — в 7 лет.
- В возрасте 4 года, 4 года 6 мес. и 5 лет вторыми по частоте встречаемости были слова из 1 слога; в 4 года они составляли 32% всех слов, в 4 года 6 мес. — 29% слов, в 5 лет — 28% слов.
- Начиная с возраста 5 лет 6 мес. вторыми по частоте встречаемости становились слова, состоящие из 3 слогов: в 5 лет 6 мес. — 27% всех слов, в 6 лет — 25%, в 6 лет 6 мес. — 26%, в 7 лет — 30%.
- С 4 лет до 5 лет 6 мес. в лексиконе детей увеличивается число слов, содержащих 3 и более слогов. С 5 лет 6 мес. до 7 лет количество слов с разным числом слогов остаётся постоянным (рис.8).

Включение в базу данных слов, содержащих артикуляционные ошибки детей, позволило провести подробное описание наиболее часто встречающихся ошибок в речи детей разного возраста [Лякso, 2008]. Термином «ошибки» обозначали все артикуляционные и грамматические отклонения от нормативного использования в русском языке, наблюдаемые в анализируемом речевом материале.

В речи детей встречаются ошибки, связанные с произнесением слов и построением фраз. Преобладающими в возрастные периоды с 4 лет до 6 лет 6 мес. являются ошибки артикуляционного плана. К 7 годам уменьшается число артикуляционных ошибок и увеличивается число ошибок во фразах (рис. 9).

Анализ артикуляционных ошибок в речи детей, записанной в разных ситуациях, показал, что большее количество детей допускает ошибки при ответе на вопросы взрослого, меньшее — в ситуации спонтанной речи.

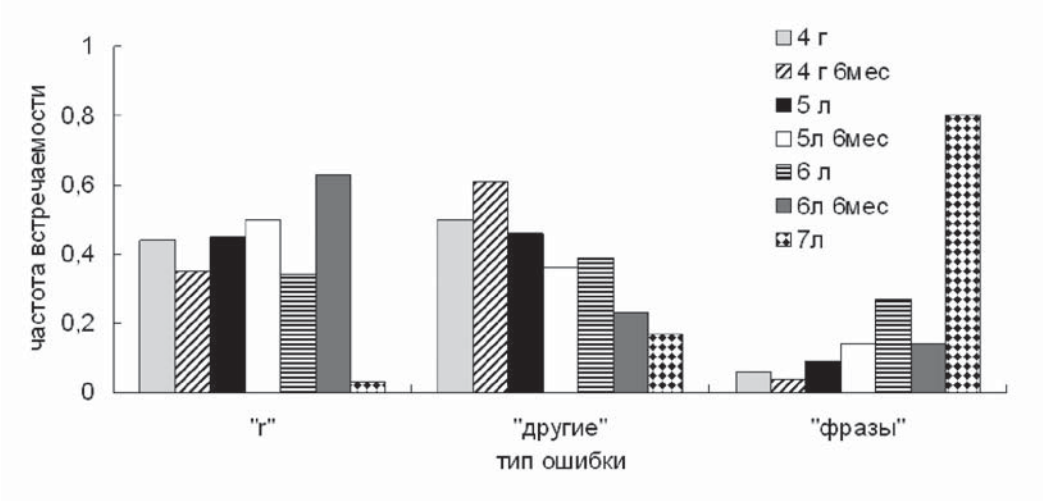


Рис. 9. Разные типы артикуляционных ошибок и ошибки во фразах в речи детей 4–7 лет (на гистограмме: светло-серый — данные для детей 4 лет, наклонная штриховка — то же для 4 лет 6мес., чёрный — 5 лет, белый — 5 лет 6 мес., горизонтальная штриховка — 6 лет, тёмно-серый — 6 лет 6 мес., ромб — 7 лет)

Встречаемость ошибок в реализации фонемы «р» (А на рис. 10) и «других» ошибок (Б) в словах, состоящих из 2 и 3 слогов, выше, чем в других словах (рис. 10), что соответствует и частотности слов в лексиконе детей.

Выявлены следующие типы ошибок реализации фонемы «р»:

- замена /р/ на /л/ (*море–моле*);
- другие варианты замены /р/: (*переодевается — пепедивается; хорошие — хаёшие*);
- пропуск /р/ с более мягким звучанием предшествующего звука *(прятали — пьятали* — замена «р» более артикуляторно простым «й»);
- пропуск /р/ *(игай-играй*);
- пропуск и замена /р/ в одном слове *(трактор — тактол*);
- пропуск слога, содержащего /р/ *(серенаду — сенаду*).

Типы ошибок индивидуальны для каждого ребёнка. В речи детей во всех анализируемых возрастных периодах значимо чаще (по сравнению с другими вариантами ошибок) встречаются замены «р» на «л» и пропуски «р». Не выявлено значимых различий в частоте встречаемости разных типов ошибок «р» в словах детей в зависимости от возраста.

Ошибки «другие» были разделены на ошибки, связанные с произнесением согласных, гласных и со слоговой структурой слова. Выявлено уменьшение количества ошибок, связанных с произнесением согласных в слове, и увеличение количества ошибок в слоговой структуре слова с увеличением возраста ребёнка. Ошибки, связанные с согласными, обусловлены пропуском или добавлением согласного в слове либо заменой другим согласным. Ошибки в слоговой структуре слова представляют собой добавление или пропуск слога. В речи детей 4 лет ошибки связаны преимущественно с заменой согласных в словах и добавлением либо пропуском слогов. В речи детей возраста 6 лет 6 мес. ошибки в словах обусловлены пропуском согласных и слогов. Возможно, это связано с ускорением темпа речи ребёнка.

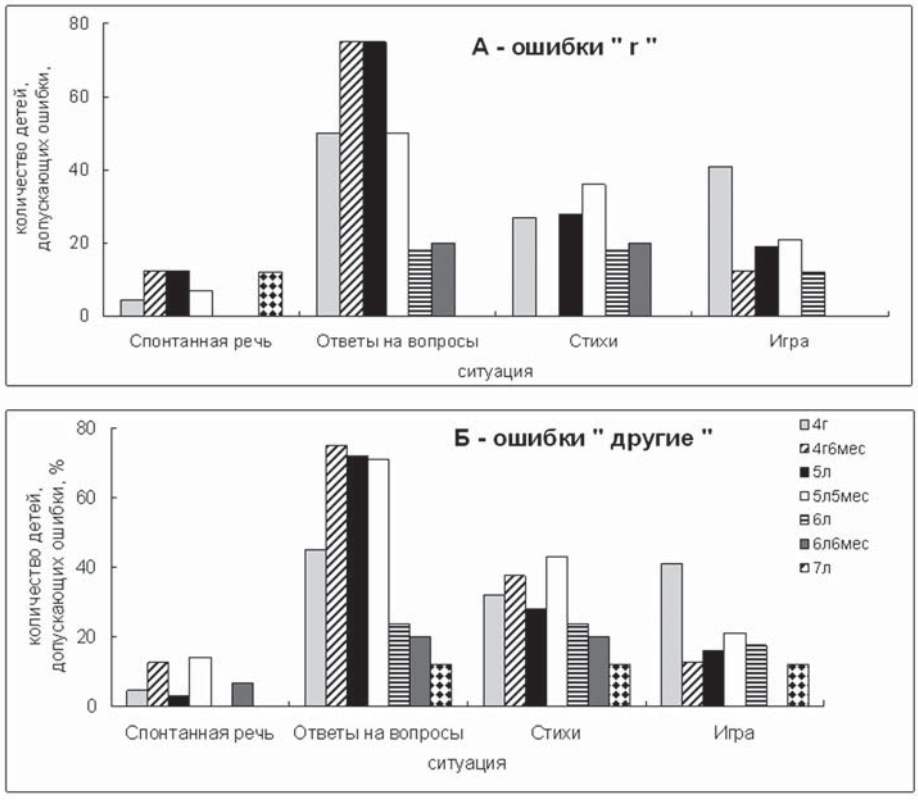


Рис. 10. Частота встречаемости слов, содержащих артикуляционные ошибки, в речи детей в разных ситуациях
(по вертикальной оси — частота встречаемости слов, содержащих артикуляционные ошибки; по горизонтальной оси — тип ситуации: спонтанная речь, ответы на вопросы взрослого, стихи или рассказ, игровая ситуация)

В основу классификации ошибок, связанных с заменой согласных, положен способ образования согласных: щелевые, смычно-щелевые, смычно-проходные, взрывные, иные (отнесены неясные варианты замен согласных).

В речи детей в возрасте 4 года наиболее часто встречаются замены щелевых согласных: /ш/ на /с/ (лошадки — ласадки); /ж/ на /с/ (книжка — книска); /ж/ на /з/ (ложится — лозится); в речи детей 6 лет 6 мес. — замены /ш/ на /с/. Установлено уменьшение разнообразия ошибок с возрастом детей и показано, что до 6 лет 6 мес. дети испытывают наибольшие сложности в произнесении слов, требующих артикуляции согласного /ш/ (рис. 11).

Таким образом, представляемые базы данных звуков и речи детей «INFANTRU» и «CHILDRU», помимо самостоятельной ценности как первой для русского языка систематической коллекции звукового и речевого материала детей в возрасте от 3 мес. до 7 лет, является хорошей основой для различных научных исследований.

Работа выполнена при поддержке фонда РГНФ (проекты № 03-06-12024 в, 06-05-12623 в) и РФФИ (проект № 09-06-00338 а).

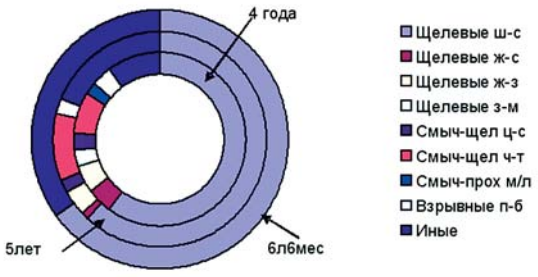


Рис.11. Различные варианты замен согласных в словах детей в возрасте 4 года, 5 лет, 6 лет 6 мес.
(разными цветами обозначены варианты замен согласных; от центра к периферии круговой диаграммы — данные для детей 4, 5 и 6 лет 6 мес. соответственно)

Литература

1. Викторов А.Б., Викторова К.О., Воронцова А.В. и др. Речевые базы данных для задач автоматического распознавания речи и верификации говорящего // Современные речевые технологии. Сборник трудов IX Сессии Российского Акустического общества. М.:ГЕОС, 1999. С. 142–145.
2. Доброва Г.Р. Эксперимент в онтолингвистике и онтолингвистический эксперимент // Проблемы онтолингвистики. Материалы международной конференции / Отв. редактор Т.А. Круглякова. СПб.: Златоуст, 2009. С. 16–21.
3. Лякso Е.Е. Артикуляционные ошибки в речи детей 4–7 лет // Сборник трудов XX сессии Российского Акустического общества. Акустика речи, медицинская и биологическая акустика. М.: Геос, 2008. Т. 3. С. 98–102.
4. Лякso Е.Е. Столярова Э.И. Специфика реализации речевых навыков 4–5-летних детей в диалоге // Психологический журнал. 2008. Т. 29. № 3. — С.48–57.
5. Лякso Е.Е., Богорад М.Ю., Гайкова Ю.С. и др. «CHILDRU»: Речевая база записей детей в возрасте от 4 до 6 лет // Сборник трудов XIX сессии Российского Акустического общества. Акустика речи и биологическая акустика. Архитектурная, строительная акустика. Шумы и вибрации. Аэроакустика. — Нижний Новгород, 2007. Т. 3. С. 83–88.
6. Лякso Е.Е., Громова А.Д., Богорад М.А. и др. База данных звуков и речи детей первых трёх лет жизни // Сборник трудов XVI сессии Российского Акустического общества. «Акустика речи. Медицинская и биологическая акустика. Архитектурная и строительная акустика. Шумы и вибрации». М.: Геос, 2005. Т. 3. С. 65–68.
7. Лякso Е.Е., Столярова Э.И., Охарева Н.Г. Речевое общение детей 4–5 лет в процессе их естественного взаимодействия // Вестн. С-Петерб. ун-та. 2008. Сер. 3. Вып. 4. С. 144–149.
8. Batliner A., Blomberg M., D'Arcy Sh., Elenius D., Giuliani D., Gerosa M., Hacke Ch., Russell M., Steidl St., Wong M. The PF STAR Children's Speech Corpus // Eurospeech — Interspeech. 2005. P. 2761–2764.
9. Bell L., Boye J., Gustafson J., Heldner M., Lindström A., Wirén M. The Swedish NICE Corpus — Spoken Dialogues Between Children and Embodied Characters in a Computer Game Scenario // Interspeech–Eurospeech. 2005. P. 2765–2768.
10. Cosi P., Pellom B.L. Italian Children's Speech Recognition for Advanced Interactive Literacy Tutors **//** Interspeech–Eurospeech. 2005. P. 2201–2204.
11. Csatári F., Bakcsi Zs., Vicsi K. A Hungarian Child Database for Speech Processing Applications // Interspeech. 1999. P. 2231–2234.
12. Galunov V.I. et. al. Wideband Speech Database for Russian // International Workshop. Speech and computer. SPB. 2002. P. 113–115.
13. Gilkerson J., Coulter K.K., Richards J.A. Transcriptional Analyses of the LENA Natural Language Corpus // LENA Foundation, Boulder, CO, LTR-06-2. 2008. Software Version: V3.1.0.
14. Giuliani D. and Gerosa M. Investigating Recognition of Children's Speech. **//** Proc. ICASSP. 2003. Vol.2. P. 137–140.

15. Hagen A., Pellom B., Cole R. Children's Speech Recognition with Application to Interactive Books and Tutors // Proc. ASRU. St. Thomas, USA. 2003.
16. Lyakso E., Frolova O. Russian Vowels System Acoustic Features Development in Ontogenesis // Interspeech. Antwerp, Belgium. 2007. P. 2309–2313.
17. Lyakso E, Bogorad M., Ostroukhov A. et. al. «INFANTRU» and «CHILDRU»: Sounds and speech databases of Russian children // Specom. Moscow. 2007. Vol. 2. P. 898–907.
18. MacWhinney B. The CHILDES project (2nd ed.). Mahwah, NJ: Lawrence Erlbaum. 1995.
19. MacWhinney B., Snow C. The Child Language Data Exchange System // Journal of Child Language. 1985. V. 12. P. 271–296.
20. Potamianos A., Shrikanth N., Sungbok L. Automatic Speech Recognition for Children // Interspeech– Eurospeech. 1997. Vol. 5. P. 2371–2374.
21. Shobaki Kh., John-Paul H., Cole R.A. The OGI kids speech corpus and recognizers // ICSLP. 2000. Vol. 4. P. 258–261.
22. Vicsi K., Roach P., Öster A., Kacic Z., Barczikay P., Sinka I. SPECO, a multimedia multilingual teaching and training system for speech handicapped children // Interspeech–Eurospeech. 1999. P. 859–862.
23. Vicsi K., Roach P., Öster A., Kacic Z., Barczikay P., Tantos A., Csatári F., Bakcsi Z., Sfakianaki A. A multimedia multilingual teaching and training system for children with speech disorders / International Journal of Speech Technology. 2000. V.3, № 3–4. P. 289–300.

Лякso Елена Евгеньевна,
доктор биологических наук, ведущий научный сотрудник кафедры Общей Физиологии Санкт-Петербургского государственного университета, руководитель группы по изучению детской речи. Lyakso@gmail.com;

Фролова Ольга Владимировна,
кандидат биологических наук, младший научный сотрудник кафедры Общей Физиологии Санкт-Петербургского государственного университета;

Громова Александра Дмитриевна,
лингвист;

Гайкова Юлия Сергеевна,
аспирантка кафедры Общей Физиологии Санкт-Петербургского государственного университета;

Куражова Анна Вадимовна,
магистрантка кафедры Общей Физиологии Санкт-Петербургского государственного университета;

Романова Ольга Дмитриевна,
магистр биологии;

Богорад Михаил Александрович,
программист;

Остроухов Александр Викторович,
фонетист, научный сотрудник Российской акустической компании «Одитек»,

Соловьёв Алексей Николаевич,
кандидат филологических наук, младший научный сотрудник кафедры Общей Физиологии Санкт-Петербургского государственного университета;

Смит Н.Ю.

Распознавание пола диктора на основе GMM-модели голоса

Ромашкин Ю.Н.,
кандидат технических наук
Петров Ю.О.

В статье рассматривается задача автоматического распознавания пола говорящего. С учётом потребности в анализе речевых сигналов относительно малой длительности предложен алгоритм решения, основанный на использовании GMM-модели голоса. Излагаются результаты экспериментальной оценки эффективности алгоритма применительно к речевым сообщениям, полученным в каналах сотовой связи стандарта GSM.

Введение

Методы автоматического распознавания речи находят всё более широкое применение в системах оказания услуг телефонной связи, туристического и гостиничного бизнеса, в технических средствах автоматических информационно-справочных служб и доступа к информации, персонализированной для каждого клиента. Они предназначены для использования произвольным абонентом, не требуют предварительного обучения и являются поэтому дикторонезависимыми. Применительно к условиям приёма речи по проводным линиям телефонной связи общего или внутрикорпоративного пользования существующие методы обеспечивают достаточно хорошую точность распознавания. Однако при использовании абонентом, например, аппаратов мобильной связи, точность распознавания может заметно снижаться, что обусловлено как воздействием помех в радиоканале, так и специфическими искажениями речи при её низкоскоростном кодировании.

Одним из возможных путей повышения эффективности методов автоматического распознавания речи может быть адаптация их параметров по гендерному признаку, т.е. в зависимости от пола говорящего. Заметные различия, например, в частоте основного тона и кратковременном спектре речи мужчин и женщин установлены в [1].

В [2] был предложен способ определения пола диктора по результатам сравнения выборочных плотностей распределения вероятностей, характеризующих значения основного тона. В целом он обеспечил хорошие результаты: вероятности правильного распознавания мужского и женского голосов составили 94,7 и 95,9 % соответственно. Однако такой подход предъявляет повышенные требования как к длительности речи (порядка 1 минуты) для минимизации статистиче-

ской погрешности оценивания выборочной плотности распределения, так и к качеству речи вследствие недостаточной помехоустойчивости оценки основного тона.

В настоящее время экспериментально доказана высокая эффективность применения GMM-метода (модель гауссовской смеси) в различных задачах речевой акустики, включая автоматическое распознавание речи, распознавание языка речевого сообщения и идентификация личности по голосу [3]. Используемые в этом методе мел-кепстральные коэффициенты обладают повышенной помехоустойчивостью и позволяют принимать достоверные решения на относительно коротких интервалах анализа речи. Поэтому интерес представляет экспериментальная оценка эффективности применения данного метода к задаче автоматического распознавания пола диктора по голосу.

1. Математическая формулировка задачи

Задача автоматического распознавания пола диктора по голосу заключается в сопоставлении некоторого речевого сообщения определённому полу диктора. Математически она может быть рассмотрена в рамках теории принятия статистических решений и сформулирована в виде проверки двух альтернативных гипотез.

Пусть задано пространство состояний, включающее две независимые последовательности *N*-мерных векторов $\vec{Y}_M(t)$ и $\vec{Y}_F(t)$ информативных признаков, характеризующие в среднем особенности мужских и женских голосов. А также образовано пространство наблюдений, состоящее из *K* записей $x_i(t) = s_i(t) + \xi_i(t),\ i = \overline{1,K},\ t = \overline{0,T}$, речевых сигналов $s_i(t)$ произвольных дикторов, принятых на фоне помех $\xi_i(t)$. Задача автоматического распознавания пола диктора по голосу состоит в установлении принадлежности каждого наблюдаемого сигнала $x_i(t)$ одному из двух возможных полов.

Переходя от реализации случайного процесса $x_i(t)$ к одноименному *N*-мерному вектору признаков $\vec{X}_i(t)$, получим следующую эквивалентную систему для проверки двух альтернативных статистических гипотез:

$$\begin{cases} H_0 : p[\vec{X}_i(t)] = p[\vec{Y}_M(t)], \\ H_1 : p[\vec{X}_i(t)] = p[\vec{Y}_F(t)]\ ,\ i = \overline{1,K}, \end{cases}$$

т.е. компоненты наблюдаемого вектора признаков принадлежат одному из двух генеральных распределений.

Используем в качестве информативных признаков кратковременные мел-кепстральные коэффициенты и применим аппроксимацию их выборочных распределений с помощью взвешенной суммы M нормальных плотностей распределения с неизвестными параметрами (GMM-модель):

$$p(\vec{X}_i \mid \lambda) = \sum_{j=1}^{M} a_j p(\vec{X}_i \mid \lambda_j),$$

где $p(\vec{X}_i \mid \lambda_j)$, $j = \overline{1, M}$, — базисные нормальные плотности распределения этих коэффициентов, a_j — вес *j-й* базисной плотности. Весовые коэффициенты имеют ограничение $\sum_{j=1}^{M} a_j = 1$. Каждая базисная плотность является *N*-мерной гауссовой функцией

$$p(\vec{X}_i \mid \lambda_j) = \frac{1}{(2\pi)^{N/2} \mid D_j \mid^{1/2}} \exp\left\{-\frac{1}{2}(\vec{X}_i - \vec{\mu}_j)^* D_j^{-1} (\vec{X}_i - \vec{\mu}_j)\right\}$$

с вектором $\vec{\mu}_j$ средних значений (размерностью *N*) и ковариационной матрицей D_j (в общем случае размерностью $N \times N$).

Параметры GMM-модели представляются в следующем виде:

$$\lambda_j = \{a_j, \vec{\mu}_j, D_j\}, \quad j = \overline{1, M}.$$

Они характеризуют индивидуальные особенности голоса каждого диктора и подлежат оцениванию по обучающей реализации речевого сигнала. Нахождение значений параметров GMM-модели голоса диктора, которые наиболее точно отражают выборочные распределения векторов признаков, осуществляется с помощью алгоритма K-средних и EM-алгоритма [4]. Сформируем таким образом средние по множеству обучающих реализаций речевых сообщений GMM-модели мужских (λ_M) и женских (λ_F) голосов.

При статистической независимости последовательности векторов признаков, наблюдаемой на интервале *T*, получим следующее выражение для логарифма функции правдоподобия:

$$P(\vec{X}_0, \vec{X}_1, ..., \vec{X}_T \mid \lambda) = -\ln \prod_{t=0}^{T} p(\vec{X}_i(t) \mid \lambda_j) = -\sum_{t=0}^{T} \ln p(\vec{X}_i(t) \mid \lambda_j).$$

Наконец, применяя критерий максимума апостериорной вероятности, решение о соответствии наблюдаемой последовательности одной из моделей λ_M или λ_F можно записать в следующем виде:

$$R[x_i(t)] = \max[P(\vec{X}_0, \vec{X}_1, ..., \vec{X}_T \mid \lambda_M), P(\vec{X}_0, \vec{X}_1, ..., \vec{X}_T \mid \lambda_F)]. \qquad (1)$$

2. Описание обучающей базы речевых сообщений

База данных речевых сообщений, использованных для создания средних GMM-моделей мужских и женских голосов, содержала цифровые записи (при частоте дискретизации 11025 Гц и 16-битном квантовании) телефонных переговоров абонентов сотовой связи стандарта GSM. Возраст абонентов составлял от 21 до 55 лет с примерно равномерным их распределением по трём возрастным группам: 21–30, 31–40 и 41–55 лет.

Для создания средних моделей λ_M и λ_F в экспериментах использовались записи (126 для мужчин и 30 для женщин) речи различных абонентов суммарной длительностью примерно 50 минут с предварительно удалёнными паузами.

В качестве информативных признаков, характеризующих индивидуальные особенности голоса абонента, использовались следующие акустические параметры речи:
— мел-кепстральные коэффициенты (*C*);
— первые производные мел-кепстральных коэффициентов (ΔC);
— вторые производные мел-кепстральных коэффициентов ($\Delta^2 C$).

3. Условия проведения экспериментов

Существующая практика применения GMM-модели в различных задачах обработки речи не даёт чётких рекомендаций о парциальных вкладах используемых параметров *C*, ΔC и $\Delta^2 C$ в общую эффективность алгоритма обработки. В большинстве исследований по умолчанию используются все три указанных информативных признака в предположении их априорной равнозначности.

В проведённых экспериментах последовательно вычислялись два варианта GMM-модели, объединяющие признаки (*C*, ΔC) и (*C*, ΔC, $\Delta^2 C$) соответственно. Парциальные вклады каждого признака оценивались в линейном приближении из условия максимизации вероятности P_D правильного распознавания пола диктора:

$$P_D = \max_{0\le\alpha\le1}[\alpha P_1(C) + (1-\alpha)P_2(\Delta C)],$$
$$P_D = \max_{0\le\beta\le1}\{\beta[\text{a } P_1(C) + (1-\beta)P_2(\Delta C)] + (1-\beta)P_3(\Delta^2 C)\}, \quad (2)$$

где P_1, P_2 и P_3 — оценки вероятности правильного распознавания, получаемые при раздельном использовании каждого из признаков.

Вычисление мел-кепстральных коэффициентов и их производных проводилось для сегментов речевых сигналов постоянной длительности 12 мс. с использованием стандартных функций среды Matlab 7. Размерность *N* соответствующих векторов равнялась 16. Матрица D_j имела диагональный вид (т.е. компоненты вектора признака считались статистически независимыми).

4. Результаты экспериментов

В экспериментах по оценке эффективности алгоритма автоматического распознавания использовались 90 тестовых записей речевых сообщений абонентов-мужчин и 30 записей женщин, полученных в канале сотовой связи стандарта GSM. При этом записи, использованные на этапах обу-

чения и тестирования алгоритма, не перекрывались как по составу абонентов, так и по времени. Длительности речевых сообщений при тестировании составляли 10 и 5 секунд (с автоматически удалёнными паузами).

Сначала выбирался порядок GMM-модели, равный 4, и для него находилось оптимальное значение весового коэффициента α, удовлетворяющее первому уравнению в (2). Результаты этих экспериментов при длительности тестовых сигналов T=10 и 5 секунд представлены графически на рис. 1 и 2 соответственно в виде зависимостей вероятности правильного распознавания пола диктора от α.

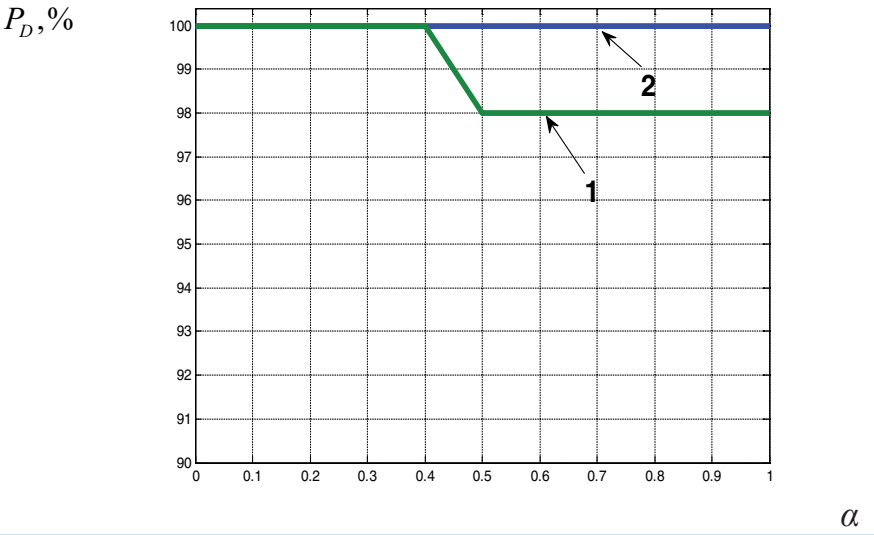


Рис. 1. Экспериментальные оценки точности распознавания пола диктора в зависимости от коэффициента α (T=10 с.): 1 — для женских голосов; 2 — для мужских голосов

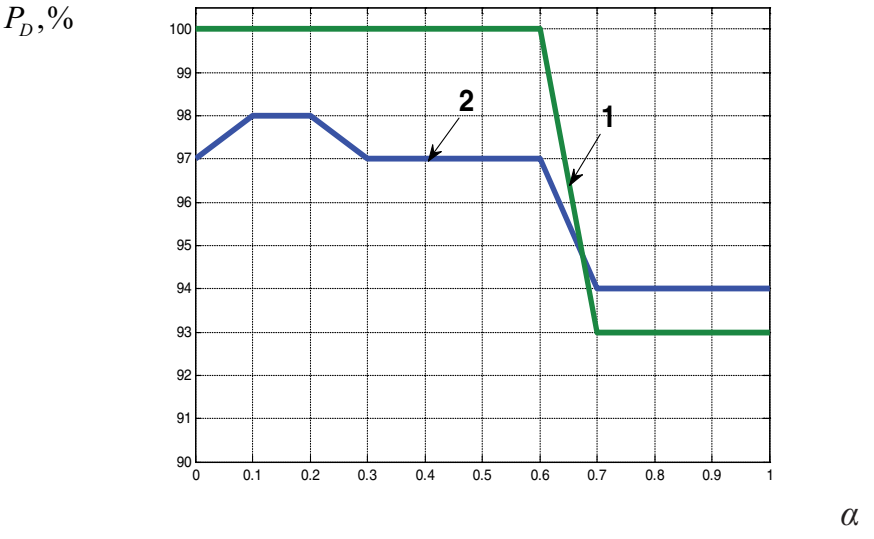


Рис. 2. Экспериментальные оценки точности распознавания пола диктора в зависимости от коэффициента α (T=5 с.): 1 — для женских голосов; 2 — для мужских голосов

Из полученных результатов следует, что при T=5–10 секунд наиболее рациональными являются значения α=0,1–0,2, т.е. относительный вклад мел-кепстральных коэффициентов в суммарную эффективность алгоритма оказался существенно меньшим, чем вклад их первых производных. Данный эффект можно объяснить тем, что в реализованном алгоритме к вектору мел-кепстральных коэффициентов не применялись известные методы нормализации [3], поэтому функции компенсации амплитудно-частотных характеристик каналов приёма и передачи речи, изменяющегося расстояния до микрофона мобильного телефона и аддитивных помех в радиоканале в этом случае переносятся на первые производные коэффициентов.

При указанных выше значениях α женские голоса распознаются алгоритмом безошибочно при T=10 и 5 с., а мужские — также безошибочно при T=10 с. и с ошибкой, равной 2 %, при T=5 с. При аппроксимации результатов испытаний биномиальным распределением доверительные интервалы полученных оценок вероятности правильного распознавания при T=5 с. и коэффициенте доверия 0,95 составили (91,5–100) % для женских голосов и (93,1–99,6) % для мужских [5]. Более узкий доверительный интервал для последних является следствием того, что тестовая выборка записей для мужских голосов оказалась в экспериментах более представительной (126 записей), чем для женских (30 записей).

Далее оптимизировалось значение весового коэффициента β в соответствии со вторым уравнением в (2) при фиксированном α=0,2. Результаты этих экспериментов представлены графически на рис. 3 и 4 в виде аналогичных зависимостей $P_D(\beta)$. Они показывают, что наиболее рациональными можно считать значения β=0,6–1,0. Однако добавление второй производной мел-кепстральных коэффициентов по времени практически не повышает эффективность распознавания, требуя при этом дополнительного времени обработки. Алгоритм по-прежнему безошибочно распознал все женские голоса при T=10 и 5 с., а также мужские

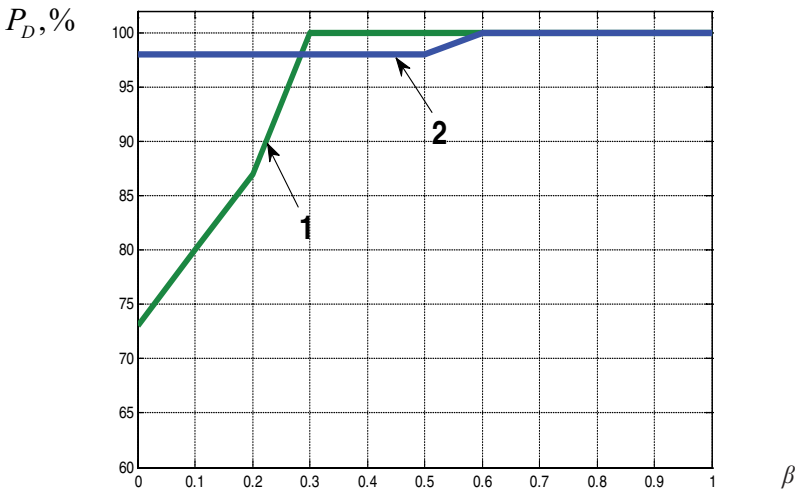


Рис. 3. Экспериментальные оценки точности распознавания пола диктора в зависимости от коэффициента β (T=10 с.): 1 — для женских голосов; 2 — для мужских голосов

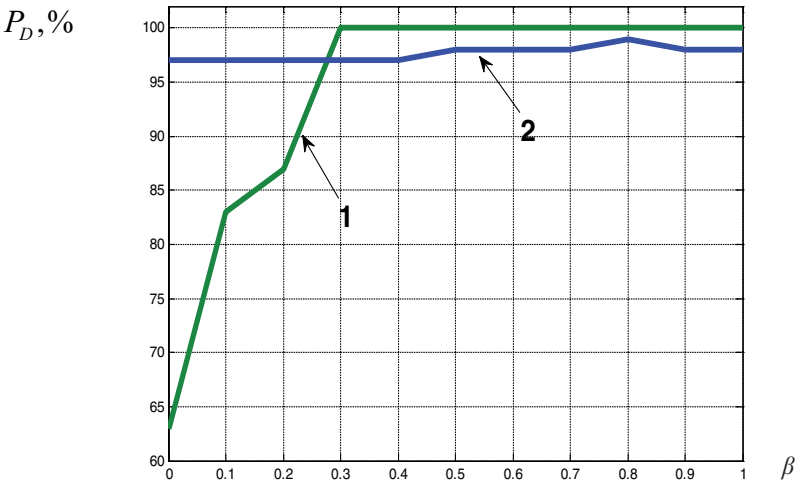


Рис. 4. Экспериментальные оценки точности распознавания пола диктора в зависимости от коэффициента β (T=5 с.): 1 — для женских голосов; 2 — для мужских голосов

при T=10 с. Ошибка распознавания мужских голосов при T=5 с. не уменьшилась и составила 2 %.

Очевидный интерес представляет поиск возможностей повышения точности распознавания при малых (T=5 с.) длительностях речевых сообщений за счёт увеличения порядка используемой GMM-модели. Результаты проведённых экспериментов (с использованием мел-кепстральных коэффициентов и их первых производных при α=0,2) показывают, что уже при увеличении порядка GMM-модели в 2 раза (до M=8) алгоритм обеспечил безошибочное распознавание пола абонентов для всех тестовых записей как с мужскими, так и женскими голосами. Доверительный интервал полученных оценок вероятности правильного распознавания мужских голосов в этом случае составил (96,7–100) % при коэффициенте доверия 0,95.

Заключение

Алгоритмы на основе GMM-модели речи могут успешно применяться для решения различных задач обработки речевой информации, в том числе автоматического распознавания пола говорящего. Такой подход наряду с повышенной эффективностью распознавания позволяет снизить требования к длительности анализируемого речевого сигнала. Полученные экспериментальные результаты показывают возможность надёжного распознавания пола при анализе коротких речевых сообщений абонентов сотовой телефонной связи стандарта GSM.

Литература

1. Михайлов В.Г., Златоустова Л.В. Измерение параметров речи. М.: Радио и связь, 1987. 168 с.

2. Сорокин В.Н., Макаров И.С. Определение пола диктора по голосу // Акуст. журнал, 2008. Т. 54. № 4. С. 659–668.

3. Benesty J., Sondhi M., Huang Y. Springer Handbook of Speech Processing, 2008. 1176 p.

4. Аграновский А.В., Леднов Д.А. Теоретические аспекты алгоритмов обработки и классификации речевых сигналов. М.: Радио и связь, 2004. 164 с.

5. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука, 1983. 416 с.

Ромашкин Юрий Николаевич
кандидат технических наук. Московский государственный институт радиотехники, электроники и автоматики (технический университет)

Петров Юрий Олегович
Государственное учреждение «Войсковая часть 35533».

Универсальная методика подготовки компонентов обучения систем распознавания речи

Викторов А.Б.,
кандидат технических наук
Грамницкий С.Г., Гордеев С.С., Ескевич М.В., Климина Е.М.

Технологию распознавания речи можно разделить на две системы: обучения и распознавания. Точность распознавания речи в определённой степени зависит от качества материала системы обучения. В данной статье описывается универсальная гибкая методика создания и тестирования компонентов системы обучения, применимая к любому языку.

Speech recognition technology might be divided into two parts — training and recognition stages. The accuracy of speech recognition depends a lot on the quality of training material. In this article we describe the flexible procedure of creating and testing the components of training system which might be applied to any language.

Введение

Основную трудность при построении систем распознавания слитной речи представляет собой не собственно создание алгоритмов распознавания на низком акустическом уровне, а построение языковых моделей на более высоком лингвистическом уровне. При этом остаётся задача качественного построения эталонов фонетических единиц языка, на которых собственно и ведётся распознавание. Следовательно, при стандартном разделении технологии распознавания речи на две относительно независимые системы обучения и распознавания, как показано на рисунке 1, от разработчиков требуется более внимательное отношение именно к системе обучения, которое выражается в более тщательном сборе и подготовке текстового и речевого корпусов [1]. Именно на основе этих корпусов происходит построение фонетического словаря, а также настройка параметров языковой модели и эталонов акустических единиц языка, которые затем используются в системе распознавания.

Таким образом, от качества корпусов обучающей системы как исходного материала, от степени структурированности содержащейся в них информации в достаточной степени зависит точность распознавания речи. В данной статье описывают-

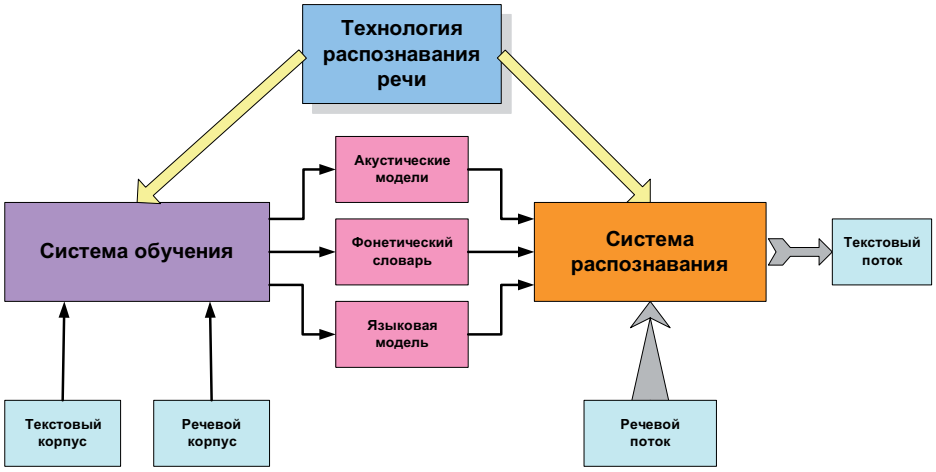


Рис. 1. Технология распознавания речи

ся апробированные методы создания универсальной гибкой системы обучения, которая пригодна для использования в технологии распознавания речи любого языка.

Конечной целью разработки данной методики было построение системы распознавания речи новостных передач. Сознательное сужение области применения системы было связано с доступностью и достаточностью материала новостной тематики для сбора и подготовки текстового и речевого корпусов, а также с объективной необходимостью применения результатов работы такой системы в коммерческой области.

Универсальность системы обучения была доказана посредством её применения на материале четырёх языков: русского, английского, немецкого и французского. Применение методов тестирования отдельных компонентов системы обучения, а также методов оценки точности распознавания речи, изменяющейся в зависимости от содержания компонентов системы обучения, обеспечило гибкость разработанной методики.

Система обучения

Система обучения представляет собой комбинацию трёх модулей, каждый из которых отвечает за подготовку одного из трёх компонентов, применяемых впоследствии в обучении системы распознавания.

Основным материалом для работы модулей системы обучения являются текстовый и речевой корпуса. В связи с этим особое внимание уделяется качеству корпусов, от которого зависит качество полученных компонентов системы обучения. Текстовый корпус применяется в работе модуля построения языковой модели и модуля построения фонетического словаря. Речевой корпус применяется в работе модуля построения акустических моделей. С помощью специально разработанных инструментов текстовый и речевой корпуса

Рис. 2. Система обучения

обрабатываются определённым образом, в результате чего создаются компоненты обучающей системы: языковая и акустическая модели и фонетический словарь.

Модуль построения языковой модели

На рис. 3 представлена схема работы модуля построения языковой модели (ЯМ).

Текстовый корпус

Основным материалом для работы данного модуля является текстовый корпус. Текстовый корпус должен отвечать следующим критериям.

1) Полнота. Текстовый корпус можно считать насыщенным в случае, если при полученном объёме корпуса прекращается резкий рост объёма новых слов.

2) Адекватность. Текстовый корпус можно считать адекватным в случае, если его тематика отвечает требованиям системы распознавания речи. В данном случае текстовый корпус должен иметь новостную тематику.

Достижение основного критерия — полноты — предполагает наличие большого количества текстов. Такой объём корпуса не позволяет проводить ручную обработку данных.

Рис. 3. Модуль построения языковой модели

Это обстоятельство явилось причиной для создания автоматизированной системы, выполняющей две основные задачи: сбора и обработки текстового корпуса. Для решения этой задачи было создано два программных продукта:

- программа для закачки текстового корпуса из Интернет-источников и очистки текста от html-тегов;
- программа для рубрикации собранных текстовых файлов, для нормализации текстового корпуса, а также для получения частотных словарей, n-граммных моделей с целью дальнейшего создания языковых моделей, а также списков ключевых слов по рубрикам с целью дальнейшей автоматизации процесса рубрикации.

Сбор текстового корпуса

С помощью специально разработанной программы был произведён сбор текстового корпуса отдельно по каждому выбранному Интернет-источнику.

Основными критериями выбора Интернет-источников являлись:

1) новостная тематика источника;
2) наличие рубрикации на сайте;
3) наличие аудио- или видеоматериалов;
4) наличие доступного (бесплатного) архива;
5) возможность обращения к архиву без использования java-скриптов.

Новостная тематика Интернет-источника обусловлена желанием построения системы распознавания речи новостных передач. Закачка текстов велась в оригинальной рубрикации сайтов. Первоначально предполагалось создать общую для четырёх языков систему авторубрикации на основе весовых функций, которая впоследствии должна была быть использована в системе распознавания для повышения эффективности работы. Однако экспериментальным путём была доказана невозможность создания такой универсальной системы авторубрикации (об этом речь пойдёт чуть позже). Критерий наличия аудио- и видеоматериалов обусловлен стремлением собрать текстовый корпус, максимально приближённый к реальной речи новостных передач. Поэтому при отборе Интернет-источников предпочтение отдавалось тем сайтам, на которых имелся подстрочник к аудио- или видеоматериалам. Однако в процессе поиска и отбора Интернет-источников выяснилось, что лишь на небольшом количестве сайтов имеется подстрочник, совпадающий с аудио- и видеоматериалом. Критерий возможности обращения к архиву без использования java-скриптов был обусловлен техническими сложностями.

В процессе работы было собрано 2 млн. 123 тыс. 441 текстовый документ с 66 одноязычных и многоязычных сайтов. Ниже приведена таблица объёма собранного корпуса по каждому из языков (таблица 1).

Таблица 1

Объём текстового корпуса

	Русский язык	Английский язык	Немецкий язык	Французский язык
Количество сайтов	25	15	16	10
Количество файлов	842 126	417 266	360 660	503 389
Количество словоформ	129 549 333	96 318 510	52 630 309	174 006 454

Рубрикация

Задачу создания универсальной системы авторубрикации текстового корпуса на основе весовых функций можно разбить на 3 подзадачи:

1) создание системы рубрик, общей для всех Интернет-источников;
2) рубрикация собранного корпуса по системе рубрик;
3) создание системы автоматической рубрикации текстового корпуса на основе списка ключевых слов.

Была создана общая для всех Интернет-источников система рубрик на разных языках. По этой системе специально разработанной программой была выполнена рубрикация на первом этапе обработки текстового корпуса. Тексты, очищенные от html-тегов и символов, были распределены по единой для всех доменов системе рубрикации. Затем были получены списки ключевых слов по каждой рубрике, после чего была проведена проверка точности рубрикации.

Поскольку оригинальная рубрикация русских сайтов выполнена в подавляющем большинстве на тематической основе, а оригинальная рубрикация зарубежных сайтов, в основном, на географической основе, был предложен многоуровневый вариант общей рубрикации. Нулевой уровень рубрикации — «Новости» — относится ко всему текстовому корпусу данного проекта. Первый уровень рубрикации («События», «Бизнес-Финансы»,

«Спорт», «Наука-Культура», «Калейдоскоп») позволяет распределить тексты независимо от типа оригинальной рубрикации сайта (тематической или же географической). Второй уровень представляет собой подробную тематическую рубрикацию для сайтов, где такую тематическую рубрикацию было возможно применить.

После получения частотных словарей по каждой из заданной системы рубрик были вычислены весовые функции рубрик. На материале тестовой выборки из корпуса, распределённого по рубрикам экспертом, выполнено тестирование, в результате которого была определена точность автоматической рубрикации на основе весовых функций. Порядок проведения тестирования:

1) из текстового корпуса, по которому были построены частотные словари, в произвольном порядке выбирается заданное количество текстов по каждой рубрике;

2) с помощью программы автоматической рубрикации тексты из выборки распределяются по рубрикам согласно полученному множеству весовых функций;

3) оценивается матрица спутывания рубрик.

Сначала данные эксперименты проводились только для русского языка по одному Интернет-домену. По каждой рубрике было случайно выбрано 30 текстов.

После проведения экспериментов были получены следующие результаты по матрице спутывания:

- количество рубрик — 7;
- средняя точность — 73%;
- максимальная ошибка спутывания — 17%;
- неизвестных документов — 6%.

Результаты получились обнадёживающие. Однако затем было проведено исследование на более обширном материале. По всему текстовому корпусу были построены весовые функции рубрик для каждого языка. По каждой рубрике из подкорпусов было отобрано по 100 текстов. Были получены матрицы спутывания для различного набора рубрик для каждого языка. В процессе проведения эксперимента количество рубрик каждый раз сокращалось путём объединения согласно матрице спутывания. В конечном итоге были получены матрицы спутывания для минимального количества общих для всех языков рубрик (трёх). Результаты этих матриц приведены в таблице 2.

Таблица 2

Результаты автоматической рубрикации

	Русский язык	**Английский язык**	**Немецкий язык**	**Французский язык**
Средняя точность (%)	48	43	60	50
Max ошибка спутывания (%)	56	51	35	50
Неизвестных документов (%)	47	35	32	47

Как показали эксперименты, выделение минимального количества рубрик (трёх) возможно лишь для немецкого языка. При этом процент спутывания и неопределённости рубрики и для немецкого языка остаётся довольно высоким. В результате проведённых исследований был сделан вывод о невозможности создания общей системы рубрикации для английского, немецкого, французского и русского языков на основе весовых функций.

Таким образом, появилась необходимость выбора и апробирования более сложной методики авторубрикации текста, нежели авторубрикация на основе весовых функций. На данный момент подобная работа ещё не проводилась.

Нормализация текстового корпуса

Нормализация текстового корпуса подразумевала:

- нормализацию орфографии, в том числе регистра символов;
- нормализацию знаков препинания;
- преобразования цифровых символов в числительные.

С помощью отдельных модулей специально разработанной программы была проведена чистка текстов по следующим этапам:

1) нормализация знаков препинания (остаются только одиночные «.» «,» «!» «?» «:» «–», отделённые от слов пробелом, точка с запятой заменяется на запятую, удаляются множественные пробелы в начале строк);
2) замена латинских одиночных букв, встретившихся в кириллическом окружении (русских словах), на кириллические и наоборот;
3) коррекция текста согласно словарю автозамен – например, исправление распространённых ошибок;
4) перевод цифровых знаков в числительные для английского и французского и удаление для русского и немецкого языков;
5) проверка по словарям на понижение регистра слов нарицательных, замену «е» на «ё». Для слов, отсутствующих в эталонных словарях, по заданным порогам на основании частоты встречаемости проводится анализ возможного регистра слова;
6) повторная коррекция текста по словарю автозамен после понижения регистра;
7) удаление повторяющихся текстов.

Комбинация и порядок модулей изменялись в зависимости от конкретной задачи нормализации текстового корпуса того или иного языка.

Построение частотных словарей

По нормализованному текстовому корпусу были построены частотные словари по рубрикам. Частотные словари были созданы для каждого языка независимо от других языков. В результате были получены следующие типы частотных словарей по каждой рубрике:

- общий частотный словарь;
- частотный словарь собственных;
- частотный словарь нарицательных.

В общий словарь были включены все слова из текстов (с учётом заданного покрытия). В словарь нарицательных и собственных — слова, разделённые в соответствии с регистром из общего словаря.

Помимо этого были получены графики зависимости объёма словаря словоформ от объёма текстового корпуса. Для каждого типа словаря строилось семейство кривых для 100%, 99%, 98%, 95% покрытия текста. Из полученных графиков с указанием области насыщения (ОН) видно, что рост новых слов при полученном объёме корпуса при покрытии 98% резко снизился. Следовательно, можно сделать вывод о достаточности набранного корпуса. На рисунке 4 представлен пример графика для общего словаря русского языка.

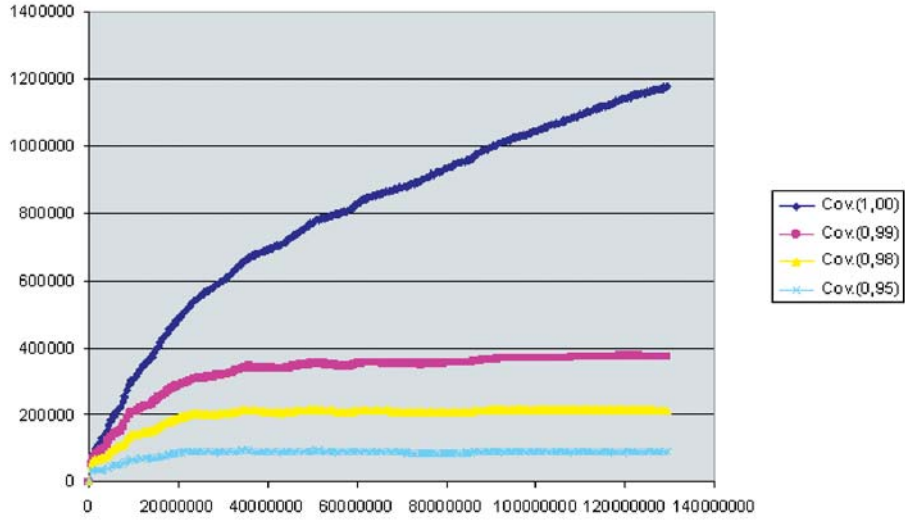


Рис. 4. График зависимости объёма общего словаря словоформ от объёма текстового корпуса; русский язык

В таблице 3 представлена информация об объёмах корпуса и словарях по четырём языкам соответственно.

Таблица 3

Покрытие

Язык	Объём словаря с покрытием 98%	Покрытие нарицательных при общем словаре с покрытием 98%	Покрытие собственных при общем словаре с покрытием 98%
Русский язык	212 950	98.67%	92.72%
Английский язык	43 119	99.09%	90.83%
Немецкий язык	223 478	99.50%	90.16%
Французский язык	67 252	98.98%	93.05%

Общие частотные словари были использованы при создании списка ключевых слов и весовых функций для системы авторубрикации, при построении n-граммных моделей для создания языковой модели, при формировании фонетического словаря.

Построение языковой модели

В настоящее время основным подходом к построению языковых моделей (ЯМ) для систем распознавания речи является использование статистических методов. При этом ЯМ в таком понимании — это просто распределение вероятности на множестве всех предложений имеющегося текстового корпуса данного языка. Для экономии памяти и увеличения быстродействия используются языковые модели, основанные на n-граммах, то есть используется явное предположение о том, что вероятность появления очередного слова зависит только от предыдущих n−1 слов. В данной системе распознавания были использованы модели со значениями n = 1, 2 и 3.

Для каждого языка файлы n-грамм были построены на последнем этапе работы специально разработанной программы на основе нормализованных текстов и полученных частотных словарей. На основании файлов n-грамм и соответствующих частотных словарей были сформированы файлы ЯМ для каждого языка. Другими словами, ЯМ являются n-граммами с соответствующими весами.

В таблице 4 приведена информация по объёму словарей, ЯМ и n-грамм для каждого языка. Отметим, что размер файлов ЯМ несколько больше, чем у n-грамм, за счёт наличия дополнительной информации о весах. Исключением является русский язык. Для него был разработан специальный алгоритм с отсечением редко встречающихся n-грамм.

Таблица 4

Данные об объёме словаря и объёме n-грамм

	Русский язык	Английский язык	Немецкий язык	Французский язык
Объём словаря с покрытием 98%	212 950	44 773	237 536	71 640
Размер файлаn-грамм (байт)	1 828 170 145	439 859 201	341 765 790	660 775 994
Размер файла ЯМ (байт)	1 009 008 146	1 034 597 103	976 488 316	1 572 253 792
Количество1-грамм	212 955	44 774	237 537	71 642
Количество2-грамм	18 973 113	6 388 776	7 467 722	8 520 606
Количество3-грамм	11 830 490	29 546 620	22 718 619	43 596 241

Оценка точности

Для анализа качества статистических языковых моделей принято использовать так называемый коэффициент ветвления (perplexity coefficient) [4,7], который можно интерпретировать как меру того, как много (в среднем) различных максимально равновероятных словоформ могут следовать за любой данной словоформой.

Для n-граммной модели коэффициент ветвления задаётся формулой:

$$Perplexity = \hat{P}(w_1, w_2, ..., w_m)^{-\frac{1}{m}} = (\prod_{t=1}^{m} P(w_t \mid w_{t-n+1}, ...w_{t-1})^{-\frac{1}{m}} = (\prod_{t=1}^{m} \frac{C(w_{t-n+1}, ..., w_t)}{C(w_{t-n+1}, ..., w_{t-1})})^{-\frac{1}{m}}$$

Это есть вероятностная оценка, приписываемая цепочке словоформ (w_1, w_2,.., w_m) языковой моделью. Здесь C — частота встречаемости данной последовательности словоформ в

обучающей выборке. Напомним, что мы рассматривали ЯМ только для n = 1, 2 и 3.

Очевидно, что коэффициент ветвления является функцией от построенной языковой модели и естественного языка (в виде текстового корпуса). Таким образом, при фиксированном языке он позволяет сравнивать различные языковые модели, а при фиксированном типе модели — оценивать сложность самих естественных языков.

После построения языковой модели был произведён подсчёт коэффициента ветвления для репрезентативной выборки файлов новостей каждого из четырёх языковых корпусов. Наличие такой оценки позволяло судить о качестве полученной ЯМ и в случае необходимости корректировать исходный материал для построения ЯМ, а именно – добирать текстовый корпус.

Модуль построения фонетического словаря

На рисунке 5 показана схема работы модуля построения фонетического словаря (ФС).

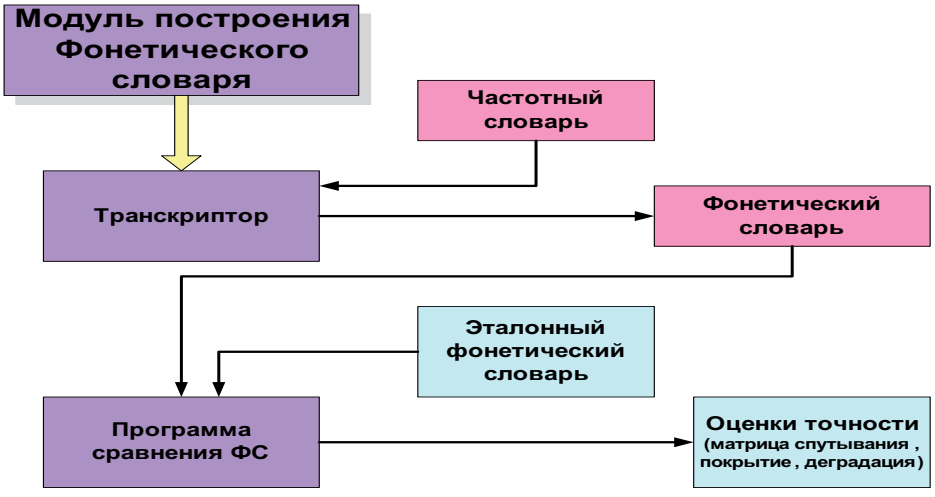


Рис. 5. Модуль построения фонетического словаря

Автоматический транскриптор

Полученные частотные словари для каждого языка были затранскрибированы специально разработанной программой — автоматическим транскриптором на основе грамматики GTT (Grammar for Text Transcription) [6]. При формировании фонетических словарей для всех языков используется фонетический алфавит SAMPA [5]. Для русского языка используется также расширенный вариант SAMPA. В этот расширенный вариант SAMPA добав-

ляются степени редукции гласных I, U, @ для заударного слога, второго и последующих предударных слогов, за исключением позиции абсолютного конца и абсолютного начала.

Автоматический транскриптор (АТ) является инструментом для преобразования письменного текста в орфографической записи на естественном языке в фонетическое представление отрезка речи, соответствующее этому тексту.

Подобные механизмы необходимы в процессе обучения систем распознавания слитной речи для установления взаимного соответствия между акустическим сигналом и фонемой как информационным элементом речи. Для этих целей можно было бы использовать словарь транскрипций, но, во-первых, словарь не может быть бесконечным и не может включать транскрипции всех слов, встречающихся в реальном речевом потоке; во-вторых, неавтоматическая генерация такого словаря — задача слишком трудоёмкая.

Построение автоматического транскриптора включает в себя три этапа:

- разработка логической структуры (абстрактного механизма) транскриптора, т.е. способов записи фонетических законов в удобной форме;
- компьютерная реализация механизмов преобразования логической структуры АТ в эффективный исполняемый код;
- разработка и запись фонетических правил для конкретного языка.

Многие существующие транскрипторы объединяют все три этапа в одном, и разработка таких АТ сводится к написанию программного кода отдельно для каждого конкретного языка. Недостатки такого подхода очевидны: реализация, модификация и поддержка таких продуктов требует от разработчика не только лингвистических знаний, но и знаний конкретного языка программирования. Кроме того, это процесс слишком трудоёмкий, и реализация АТ для каждого языка, как правило, предполагает разработку совершенно нового программного компонента.

Другой, более эффективный, подход сводится к разделению процесса написания правил транскрипции (для этого используются различные формальные языки) и реализации общего программного модуля (собственно языково-независимого АТ).

Для того чтобы обеспечить возможность написания правил транскрипции без изменения кода, нами был разработан язык описания правил транскрипции. Он не ориентирован на обработку текстов на каком-либо конкретном языке, т. е. не имеет каких-либо предопределённых классов звуковых сегментов и пр. Используемые им структуры данных могут использоваться для представления элементов звуковой системы любого языка.

Реализация данного языка представляет собой программу-интерпретатор, считывающую правила транскрипции, преобразующую их в более эффективное представление и применяющую их к входному тексту на естественном языке.

Большинство известных транскрипторов на основе правил используют собственные языки, включающие в себя ограниченный набор функций. Запись правил на таких языках формализована в той или иной степени и представляет собой системы замены цепочек входных символов на транскрипционные знаки. Такой язык обладает ограниченной функциональностью, практически нерасширяем и пригоден для применения исключительно в транскрипционном модуле.

Разработанный язык описания правил представляет собой формальную грамматику (порождающую контекстно-свободную грамматику типа AGFL (Affix Grammars over a Finite Lattice) [6]. Программа-интерпретатор преобразует грамматику, написанную для конкретного языка, в наиболее эффективную форму — разновидность конечного автомата, что обеспечивает высокие показатели быстродействия.

Формальные грамматики являются мощным средством разработки лингвистических компонентов практически любого уровня: морфологического, синтаксического и др. Разработка и применение нами формальной грамматики GTT (Grammar for Text Transcription) доказывает эффективность использования таких средств для решения задач автоматической транскрипции текста. Более того, модификация этой грамматики может применяться и в других (в том числе указанных выше) лингвистических модулях.

Разработка транскриптора проводится в четыре этапа:

1) реализация транскрибирования изолированных слов по правилам литературной нормы;
2) реализация транскрибирования слитной речи по правилам литературной нормы;
3) реализация транскрибирования изолированных слов разговорной речи;
4) реализация транскрибирования слитной разговорной речи.

В результате создания и применения транскриптора на основе формальной грамматики GTT были разработаны правила транскрибирования слов и предложений. Точность транскрипции доходит до 99%, что является отличным результатом, при этом количество правил значительно меньше по сравнению с существующими аналогами.

К несомненным достоинствам данного продукта следует отнести:

• независимость лингвистической и программной части, благодаря чему:
1) правила могут разрабатывать лингвисты, не знающие языка программирования;
2) разработка правил для новых языков и изменение существующих правил не требует изменений в коде и, соответственно, является задачей намного более простой;
• грамматика GTT обладает преимуществом по сравнению со многими языками для записи фонетических правил, поскольку:
1) несмотря на то что грамматика GTT является новым продуктом, она разработана в соответствии с уже существующими принципами, использует традиционные структуры данных, так что освоение грамматики для профессионального лингвиста не составляет труда;
2) грамматика создана с учётом особенностей фонетического анализа, но может быть легко расширена для решения задач и в других областях лингвистического анализа;
3) структура типов и формат правил грамматики позволяют наиболее точно и сжато представлять правила транскрипции, за счёт чего значительно уменьшается их количество и упрощается задача разработчика;
4) в грамматике предусмотрена возможность вариативной транскрипции, вследствие чего увеличивается точность транскрибирования;
5) грамматика поддерживает транскрипцию не только изолированных слов, но и предложений;

• программа-интерпретатор учитывает как потребности лингвиста-разработчика, так и системные требования:
1) в программе предусмотрен отладочный режим и функция сравнения транскрипций, благодаря чему разработчик может наиболее эффективно оценивать результаты работы АТ;
2) правила грамматики переписываются в эффективный код, что увеличивает быстродействие АТ.

Оценка точности

Проверка транскрипции в ФС осуществлялась в несколько этапов полуавтоматическим способом, то есть ручная проверка чередовалась с автоматической проверкой специально разработанной программой, которая проводила статистический анализ ошибок и позволяла оценить точность новой версии автоматического транскриптора.

Цель ручной поверки состояла в создании эталонного файла транскрипции, проверенной экспертом-фонетистом. Ручная проверка транскрипции осуществлялась с помощью специальной программы-редактора, где фонетист для каждого слова проставлял статусы, характеризующие верность или же ошибочность транскрипции с указанием типа ошибки.

Таким образом, можно было оценить результаты работы автоматического транскриптора на основании проверки транскрипции, выполненной экспертом.

Автоматический транскриптор может быть охарактеризован с помощью следующих параметров: точность транскрипции, её избыточность, а также сложность грамматики.

Точность — это отношение (в процентах) количества правильно сгенерированных транскрипций к числу транскрипций.

Избыточность — это отношение (в процентах) количества всех сгенерированных транскрипций к числу входных слов (100% соответствует «нулевой» избыточности, то есть для каждого слова одна, и только одна транскрипция).

Сложность грамматики — это количество используемых правил.

В конечном итоге точность транскрипции зависит как от сложности фонетических правил входного языка, так и от насыщенности подключаемых словарей. Таким образом, появилась возможность, проанализировав имеющиеся ошибки автоматического транскриптора, понять причину несовершенства ФС и устранить её.

Модуль построения акустических моделей

На рисунке 6 показана схема работы модуля построения акустических моделей АМ.

Основным материалом для модуля построения АМ был речевой корпус.

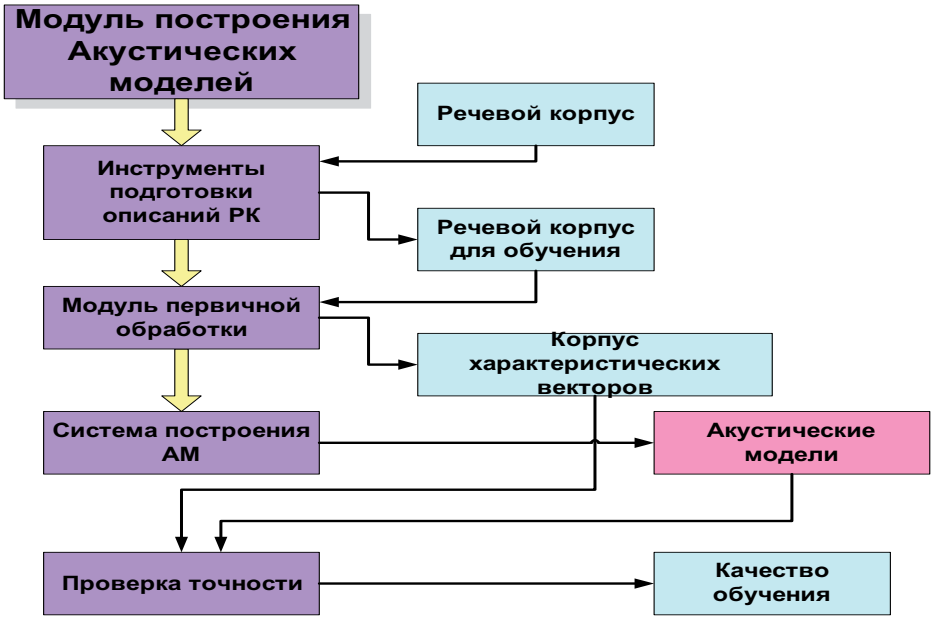


Рис. 6. Модуль построения акустических моделей

Речевой корпус

Из имеющихся речевых баз данных был создан речевой корпус (РК) для каждого языка.

Таблица 5

Объём речевого корпуса

	Русский язык	Немецкий язык	Английский язык	Французский язык
Общая продолжительность (часов)	> 200	> 20	> 20	> 20
Количество дикторов	3280	4000	4000	5000

Все речевые базы были объединены, а орфографические подстрочники к этим базам были унифицированы, для чего была создана система обозначений и правил. Например, начало предложения пишется со строчной буквы, кроме имён собственных и имён существительных в немецком языке; из знаков препинания сохраняются только точки и запятые; специальным образом маркируются неречевые акустические события, неправильно или нечётко произнесённые слова и предложения и т.п.

На основании унифицированного подстрочника с помощью инструмента обработки текстового корпуса были построены частотные словари РК, которые были преобразованы в фонетические словари автоматическим транскриптором. Кроме того, на основании этих частотных словарей были получены языковые модели РК.

В конечном итоге подготовленные описания речевых баз, предназначенных для обучения, представляли собой набор файлов в текстовом формате:

- фонетический алфавит;
- фонетический словарь;
- файл орфографического подстрочника;
- языковая модель РК.

Одновременно были подготовлены речевые базы, предназначенные для тестирования системы распознавания, которые содержат три типа файлов: файл аудиозаписи, файл орфографической записи и файл временной привязки орфографической записи к аудиофайлу.

Первичная обработка заключалась в преобразовании речевого сигнала в последовательность характеристических векторов. В качестве признаков мы использовали mel-частотные кепстральные коэффициенты с их первыми и вторыми производными.

Система построения АМ основана на «скрытых марковских моделях» (СММ) [2,3].

Акустические модели строились для таких акустических единиц, как фоны, дифоны и трифоны. В качестве акустических моделей мы использовали многокомпонентные непрерывные СММ с Гауссовой функцией распределения вероятностей появления характеристических векторов.

Проверка точности построенных АМ осуществлялась путём распознавания тестовой речевой базы данных с помощью полученных АМ только на акустическом уровне, без использования знаний о ЯМ. При таком подходе случайное событие появления каждой акустической единицы в любой момент времени имеет равновероятное распределение.

Тестирование работы системы распознавания речи

Оценка точности работы системы распознавания речи проводилась по схеме, указанной на рисунке 7.

Поступающий акустический сигнал сначала проходит этап параметрического представления — такой же, как при построении АМ. Полученные характеристические вектора анализируются классификатором, в результате чего происходит сегментация входного потока на такие классы, как речь, шум, пауза и музыка. В дальнейшем декодер речи работает только с сегментами, которые помечены как речь. На этапе декодирования речи используются все компоненты: акустические модели, фонетический словарь, языковая модель.

Тестирование системы распознавания производилось в несколько этапов. На каждом этапе использовались различные сочетания версий компонентов обучающей системы: акустические модели, полученные на основе различных речевых корпусов; фонетические словари, полученные различными версиями автоматического транскриптора, с использованием обычного и расширенного варианта SAMPA; языковые модели, построенные с учётом редких n-грамм и без их учёта, построенные на текстовых корпусах разного объёма. Комбинирование различных версий используемых компонентов позволяло отслеживать степень влияния того или иного компонента на результаты тестирования и находить пути улучшения точности распознавания посредством

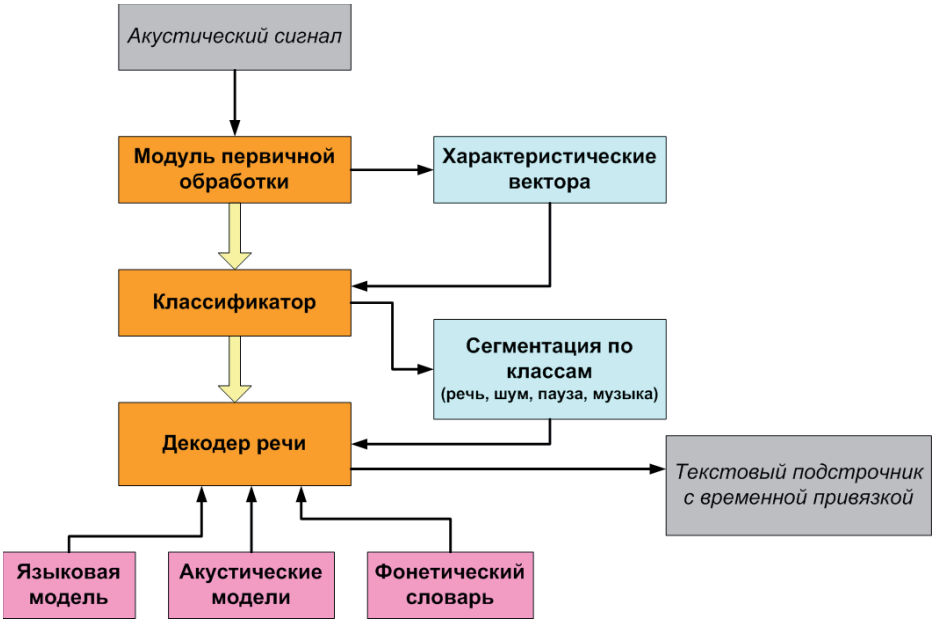


Рис. 7. Оценка точности распознавания речи

усовершенствования отдельных компонентов. В результате этого была достигнута точность распознавания 60–70% в зависимости от качества звуковых файлов.

Заключение

В результате проделанной работы была создана универсальная гибкая система обучения, в которой используются многофункциональные инструменты обработки данных, а система тестирования обеспечивает определённую гибкость процесса повышения эффективности распознавания речи. Таким образом, описанная методика применима к любому языку и позволяет повышать точность распознавания речи путём совершенствования определённого компонента системы обучения.

В данный момент ведётся работа над повышением точности распознавания для упомянутых языков (русского, английского, немецкого и французского), а также над привлечением материалов других языков для дальнейшего апробирования работы данной технологии.

Литература

1. *Кривнова О.Ф.* Речевой корпус на новом технологическом витке // Речевые технологии. М., 2008.
2. *Марков А.А.* Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь // Известия Академии наук. СПб. VI. Т.7. 1913. № 3. С. 153–162.

3. *Марков А.А.* Об одном применении статистического метода. Доклад в Академии наук от 17 февраля 1916 года.
4. *Bahl L.R., Baker J.K., Jelinek F., Mercer R.L.* Perplexity — A measure of the difficulty of speech recognition tasks. // J. Acoust. Soc. Amer. Vol.62. P.S63. 1977. Suppl. no.1.
5. http://www.phon.ucl.ac.uk/home/sampa/
6. http://www.agfl.cs.ru.nl/
7. *Wųlfel M., McDonough J.* Distant Speech Recognition. 2009.

Викторов Андрей Борисович
кандидат технических наук,
заместитель генерального директора по науке ООО «ОДИТЕК».
В 1985 году окончил Политехнический институт
(Физико-механический факультет, кафедра Прикладной математики).
Опыт работы в области речевых технологий с 1985 года в НПО «Дальняя связь».

Грамницкий Сергей Николаевич
руководитель проекта ООО «ОДИТЕК», окончил ЛЭТИ,
опыт работы в области речевых технологий с 2000 года

Гордеев Станислав Сергеевич
программист ООО «ОДИТЕК», окончил СПбГУ
(Филологический факультет,
кафедра Теоретической и прикладной лингвистики),
опыт работы в области речевых технологий с 2006 года

Ескевич Мария Владимировна
лингвист ООО «ОДИТЕК», окончила СПбГУ
(Филологический факультет, кафедра Теоретической и прикладной лингвистики),
опыт работы в области речевых технологий с 2004 года

Климина Екатерина Михайловна
лингвист ООО «ОДИТЕК», в 2006 году окончила СПбГУ
(Восточный факультет, кафедра Индийской филологии),
опыт работы в области корпусной лингвистики с 2006 года.

Моделирование речевой просодии: временной компонент выделительного акцента в английском языке

Филясова Ю.А.

В предлагаемой статье рассматривается реализация акцентного выделения (АВ) в английском языке на участках гласных в составе односложных слов, выделенных из кратких диалогов, в чтении одного диктора. Автор обнаруживает различия в длительности гласных в зависимости от позиции акцентированного слова во фразе и глухости/звонкости конечного согласного слова и делает вывод о том, что выделительный акцент в английском языке реализуется в условиях взаимодействия просодических и сегментных факторов. С целью возможного практического применения полученных данных выделен процент прироста длительности гласных, ответственный за реализацию АВ в заданных условиях.

Введение

Речевые технологии как одно из направлений информационных технологий занимается проблемами общения человека с компьютером (или человека с человеком посредством компьютера) на основе использования языка в его звуковой форме. Речевые технологии используются в создании систем автоматического синтеза и распознавания речи; при построении современных средств речевого общения со сложными техническими устройствами, занимающими всё более значительное место в разных сферах жизни общества [1], в частности, при создании человеко-машинных интерфейсов с устным вводом/выводом информации, при организации информационно-справочной службы, в обучении иностранному языку (автоматические фонетические тренажёры) и т.д.

Современные программные и технические средства позволяют моделировать естественные процессы порождения и восприятия речи. При этом необходимо помнить, что эффективность моделирования речевого сигнала и

решения других прикладных задач определяется полнотой использования фонетических сведений, полученных при изучении свойств естественной речи. Особое значение для речевых технологий представляют сведения о просодической организации устной речи, которые ответственны за естественность звучания синтезированной речи. Так, например, формирование фразовых интонационно-просодических показателей (типов интонации, особых подчёркиваний слов и т.д.) является одной из основных задач на этапе лингвистической обработки в системах «Текст — Речь» [2]. Изучение просодических явлений осложняется их внутренней взаимозависимостью на различных уровнях преобразования (лексическом, фонетическом, прагматическом). Каждый из просодических параметров должен извлекаться в результате довольно сложного анализа, который сам по себе варьируется в зависимости от степени продвинутости инструментария, речевых технологий и т.д. [3].

В данной статье приводятся некоторые данные исследования длительности гласных в условиях акцентного выделения[1]. АВ является языковым средством подчёркивания особо важных элементов высказывания. Местоположение его в высказывании не фиксировано, поскольку обусловлено конкретными коммуникативными задачами.

Длительность, по данным современных исследований, является важной супрасегментной характеристикой выделения основных и фоновых (второстепенных) смысловых фрагментов текста. Сегментная длительность служит одним из признаков эмфатического или противопоставительного ударения [5]. Поскольку во многих языках длительность выполняет функцию различения фонем, долгое время её не рассматривали в качестве одного из главных просодических параметров. Одним из первых, кто начал рассматривать длительность как один из наиболее информативных компонентов просодии, был Д. Фрай [6]. Он делает вывод о том, что увеличение длительности в английском языке — более эффективное средство выделения, по сравнению с увеличением интенсивности.

В изучении акустических параметров АВ в английском языке длительность приобретает особое значение, поскольку определённым образом участвует в сегментной организации речи. Ещё в XIX веке в работах по фонетике учёные писали о сокращении длительности гласных в контексте перед глухими согласными [7]. Современные исследователи отмечают регулярный характер данного явления в английском языке [8; 9, с. 58].

Существует также универсальная фонетическая закономерность — увеличение длительности сегментов к концу коммуникативной единицы, — которая характеризует любое устное высказывание и не зависит от сегментного состава единиц, входящих в это высказывание. Так, например, Д. Клатт определил, что в английском языке длительность гласных в области перед паузой увеличивается на 30% [10].

Кроме того, особенностью английского языка является строгая ритмическая организация устной речи («stress-timing»). Одним из первых её отметил Д. Аберкромби [11]. Ударные слоги реализуются через примерно равные промежутки времени, независимо от количества безударных слогов между ними. Длительность при этом выступает в качестве средства для оформления равных ритмических отрезков высказывания.

С учётом указанных функций темпорального компонента в английском языке, сведения об изменениях длительности гласных английского языка в условиях акцентного выделения могут внести значительный вклад в лингвистические основания речевых технологий.

---

[1] Мы используем термин «акцентное выделение» (далее — АВ) вслед за Т.М. Николаевой, см. [4].

Эксперимент

В данной статье описываются результаты исследования, ***задача*** которого состояла в том, чтобы выявить особенности реализации длительности гласных при АВ в условиях влияния таких фонетических факторов, как: 1) сегментный (правый) консонантный контекст; 2) фразовая позиция.

Материалом исследования послужила запись микродиалогов в чтении носителей британского варианта английского языка, которая выполнялась в Лаборатории экспериментальной фонетики (ЛЭФ) им. акад. Л.В. Щербы СПбГУ. Для анализа был отобран материал одного диктора. Исследованию подлежали гласные в словах структуры CVC[2] в разных фразовых позициях (табл. 1). Длительность гласных в неакцентных позициях выступает в качестве объекта для сравнения длительности гласных в акцентных позициях. Нейтральной позицией, т.н. точкой отсчёта, стала «*IA*-»[3] (на рис. 3 отмечена овалом). Для обработки материала использовались программы Wave Assistant v/2.00 (Центр Речевых Технологий (ЦРТ), Санкт-Петербург, 1998), Excel, Statistica v/5.00. Слова условно делятся на две группы: со звонким (V+) и глухим (V–) конечным согласным, например, bag — back, cad — cat. Количество реализаций гласных в каждой заданной позиции насчитывает 30 ед., как в контексте перед глухими, так и в контексте перед звонкими согласными. Общее количество исследованных гласных составляет 360 ед.[4]

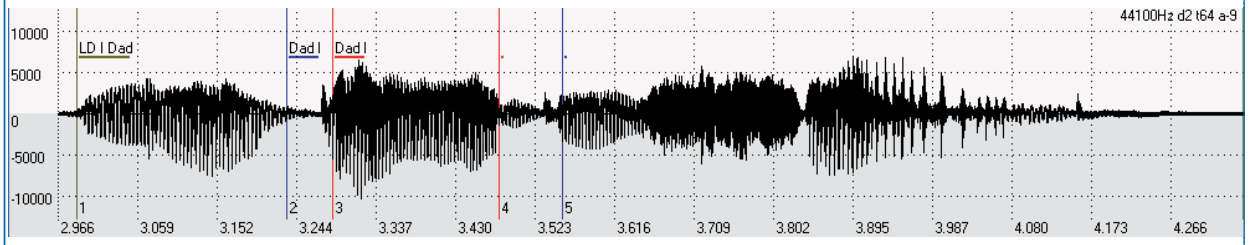
Таблица 1

Фразовые позиции для слова

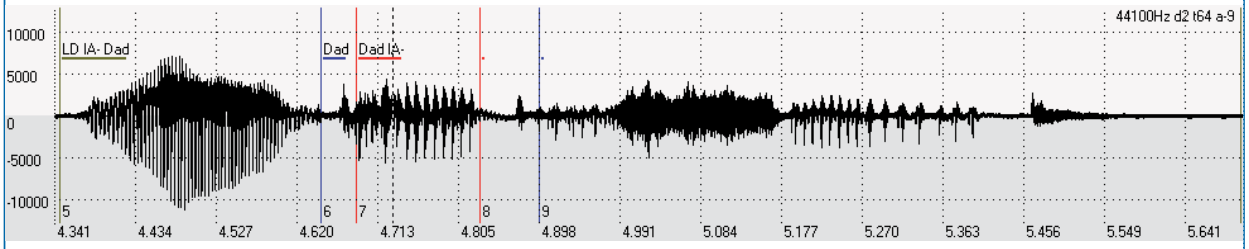
№	Условное обозначение позиции	Позиция слова во фразе относительно выделенности	Образцы микродиалогов на примере слова «*bag*»
1	***«I»*** *(Initial)*	Слово под ударением	*The «bag» was found.*
2	***«IA–»*** *(Initial Accented)*	Слово в позиции после выделенного слова	*A: Whose bag was found? Was your bag found?* *B: No,* ***your*** *«bag» was found.*
3	***«IA+»*** *(Initial Accented)*	Слово с выделительным ударением	A: What was found? A sack? *B: No, a «****bag****» was found.*
4	***«F»*** *(Final)*	Слово под синтагматическим ударением	*This is a «bag».*
5	***«FA–»*** *(Final Accented)*	Слово в позиции после выделенного слова	A: What a small sack you bought! *B: But this is not a sack, this is a bag.* *A: Since I took it for a sack,* *it is a* ***very big*** *«bag».*
6	***«FA+»*** *(Final Accented)*	Слово с выделительным ударением	A: What a small sack you bought! *B: But this is not a sack, this is a «****bag****».*

[2] CVC: Consonant — Vowel — Consonant (Согласный — Гласный — Согласный).
[3] Мы считаем, что гласный в безударной позиции в начале фразы менее всего подвержен влиянию различных просодических факторов.
[4] В данной статье представлены результаты исследования длительности гласных в составе утвердительных фраз.

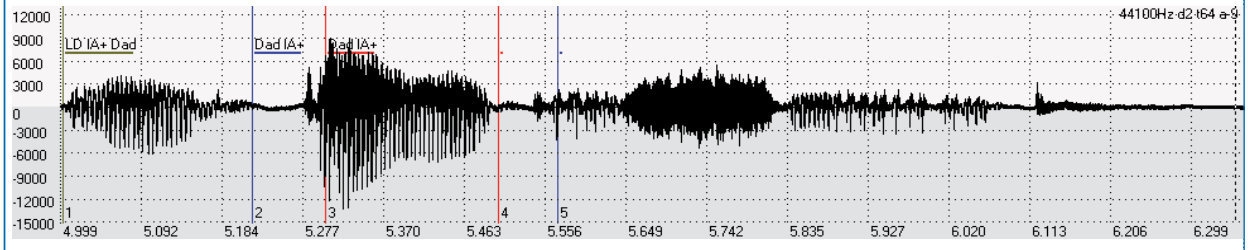
На рис. 1 и 2 представлены примеры фраз, содержащие отобранные для исследования слова. Каждый рисунок демонстрирует слово в одной из заданных фразовых позиций. Для примеров выбраны два слова, контрастные по сегментному составу: «*Dad*» *vs*. «*tap*». Оба слова имеют одинаковый центральный элемент — открытый гласный переднего ряда /æ/, но в одном случае перед глухим, а в другом — перед звонким согласным. Изменения длительности гласного можно проследить не только от начала к концу фразы (фразово-позиционные изменения), но и в одинаковых фразовых позициях перед разными согласными (сегментные различия).



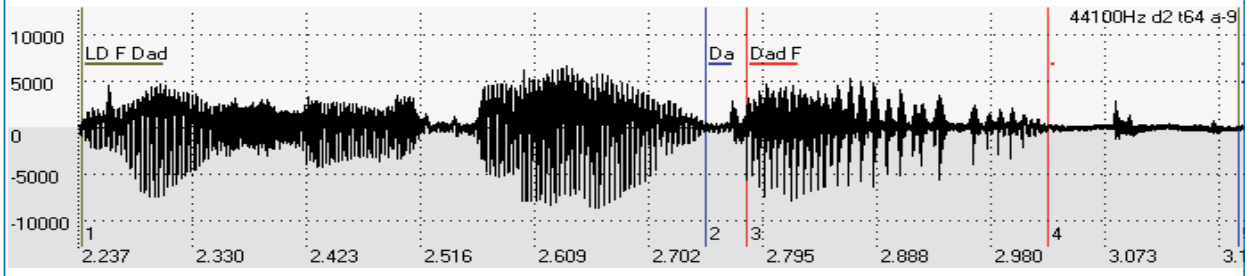
***A*: (Позиция «*I*»)** *My «Dad» is sad*[5]. (/æ/=194 мс)



***B*: (Позиция «*IA–*») *Your «Dad» is sad*.** (/æ/=143 мс)

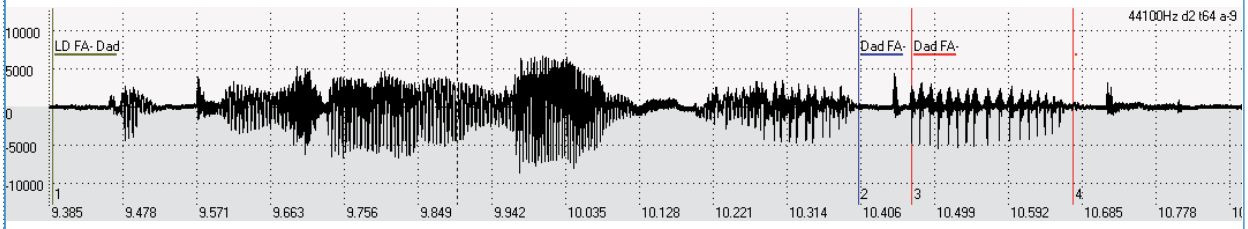


***C*: (Позиция «*IA+*») *My «Dad» is sad*.** (/æ/=199 мс)

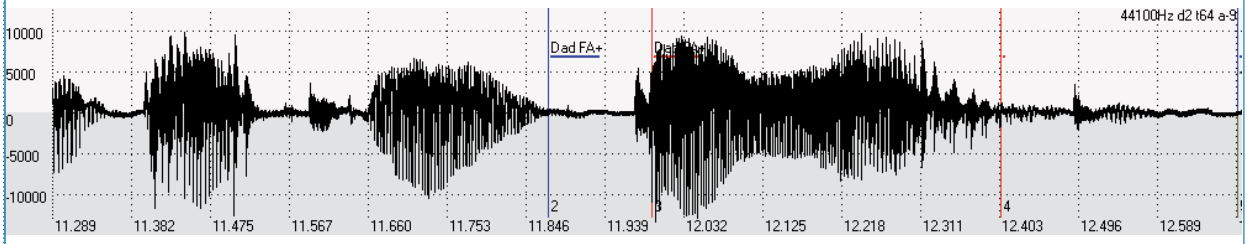


***D*: (Позиция «*F*»)** *This is my «Dad»*. (/æ/=245 мс)

[5] Слова для исследования взяты в кавычки. «Жирным» шрифтом отмечены слова с АВ. Слова в кавычках «жирным» шрифтом были отобраны для исследования и произнесены в составе диалогов с АВ.



E*: (Позиция «*FA–*»)** *Because I* ***love *my «Dad».* (/æ/=203 мс)



F*: (Позиция «*FA+*»)** (*But I'm telling you) about my «*Dad****»*. (/æ/=409 мс)

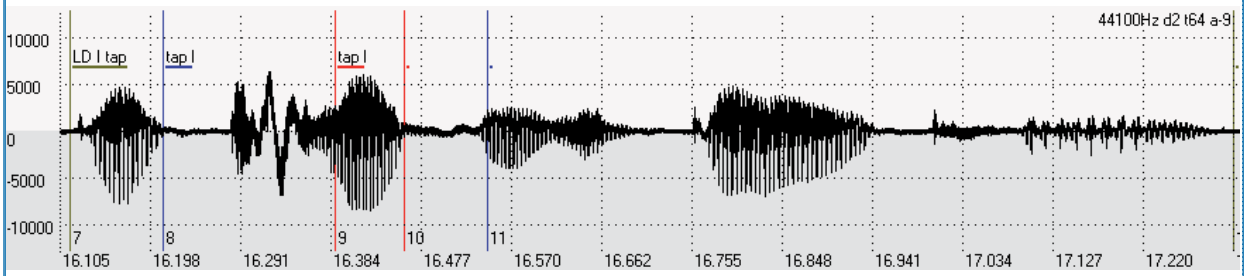
Рис. 1. Длительность гласного / æ / в заданных фразовых позициях в составе слова «Dad» [6]

На примере слова «Dad» видно, что длительность гласного в конечных позициях («*F*», «*FA–*», «*FA+*») больше, чем в начальных («*I*», «*IA–*», «*IA+*»). Примеры демонстрируют проявление общеизвестной фонетической закономерности увеличения длительности к концу высказывания. Наибольшая длительность наблюдается в конечной акцентной позиции «*FA+*» (409 мс). Следует отдельно подчеркнуть, что длительность гласных в словах с АВ в конце фразы на 184 мс больше, чем под нейтральным синтагматическим ударением без АВ (409 мс в позиции «FA+» vs. 225 мс в позиции «F»).

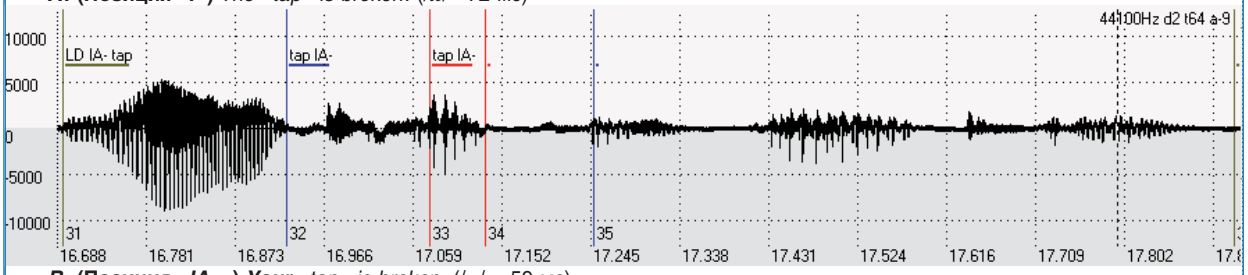
В начальной акцентной позиции «*IA+*» длительность гласного /æ/ не намного больше, чем в позиции без АВ «*I*» (199 мс vs. 194 мс), и является наибольшей только среди начальных позиций. Таким образом, длительность гласных при АВ зависит от позиции во фразе. Эта зависимость проявляется и на длительности гласных в сегментном контексте перед глухим согласным (см. рис. 2).

Сравнение длительности гласных в одинаковых позициях, но с разным правым консонантным контекстом (рис. 1 vs. рис. 2) показывает, что в слове «*tap*» длительность гласных перед глухим согласным во всех заданных позициях значительно меньше, чем в слове «*Dad*». Позиции с АВ не являются исключением. Так, например, в позиции «*IA+*»: 199 мс в слове «*Dad*» vs. 90 мс в слове «*tap*». В позиции «*FA+*»: 409 мс в слове «*Dad*» vs. 116 мс в слове «*tap*».

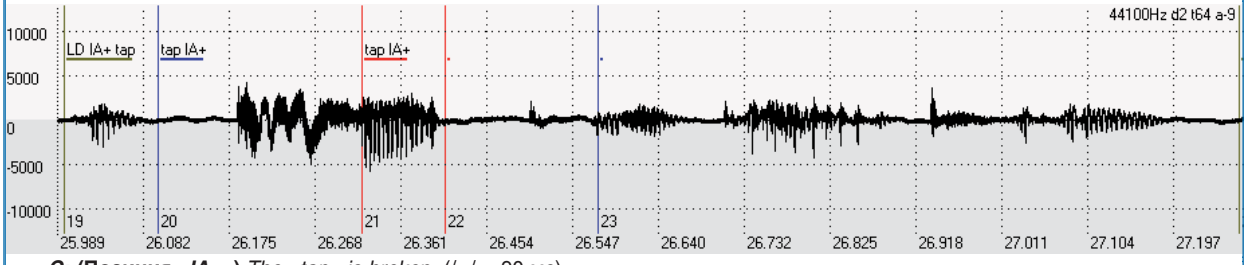
[6] Для наглядного представления фраз были заданы параметры, одинаковые для всех примеров: частота дискретизации 44100 Гц, растяжение по оси OX — 64, растяжение по оси OY — 9. Синими метками обозначены слова, красными метками — гласные, жёлтыми метками указаны границы фраз. При этом частота дискретизации понимается как частота взятия отсчётов непрерывного во времени сигнала при его дискретизации аналого-цифровым преобразователем [12].



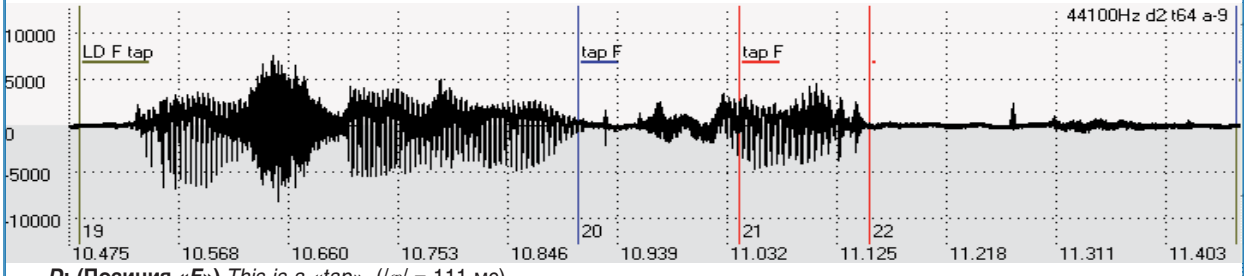
***A*: (Позиция «*I*»)** *The «tap» is broken.* (/æ/= 72 мс)



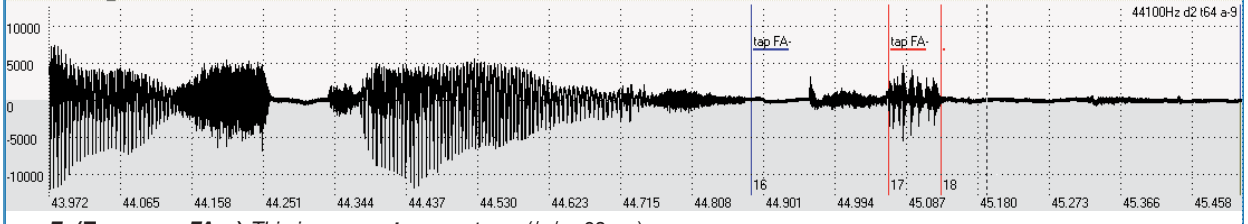
B*: (Позиция «*IA–*») *Your «tap» is broken. (/æ/ = 59 мс)



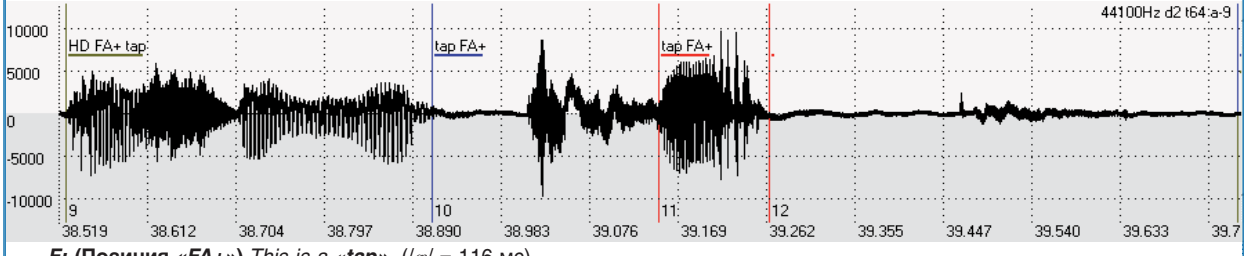
***C*: (Позиция «*IA+*»)** *The «tap» is broken.* (/æ/ = 90 мс)



***D*: (Позиция «*F*»)** *This is a «tap».* (/æ/ = 111 мс)



***E*: (Позиция «*FA–*»)** *This is a very **strange** «tap».* (/æ/ = 68 мс)



***F*: (Позиция «*FA+*»)** *This is a «**tap**».* (/æ/ = 116 мс)

Рис. 2. Длительность гласного / æ / в заданных фразовых позициях в составе слова «tap»

Поведение длительности гласных на примере указанных выше двух слов вполне соотносится со средними показателями длительности, полученными на материале 360 гласных.

Рассмотрим фразово-позиционные изменения длительности гласных в контексте перед звонкими и глухими согласными.

При больших значениях длительности гласных перед звонкими[7] согласными увеличение показателей длительности гласных наблюдается от начала фразы к её концу, как в контексте перед звонкими, так и в контексте перед глухими согласными (см. рис. 3). Так, перед звонкими согласными средняя длительность гласных изменяется от 140 мс (в начальной безударной позиции «*IA–*») до 330 мс (в конечной акцентной позиции «*FA+*»); перед глухими — от 90 мс (в начальных ударной «*I*» и безударной «*IA–*» позициях) до 170 мс (в конечной акцентной позиции «*FA+*»).

Длительность гласных в начальной акцентной позиции «*IA+*» в обоих сегментных контекстах меньше длительности гласных в конечной акцентной позиции «*FA+*» (на 100 мс в контексте перед звонкими согласными и на 50 мс – перед глухими).

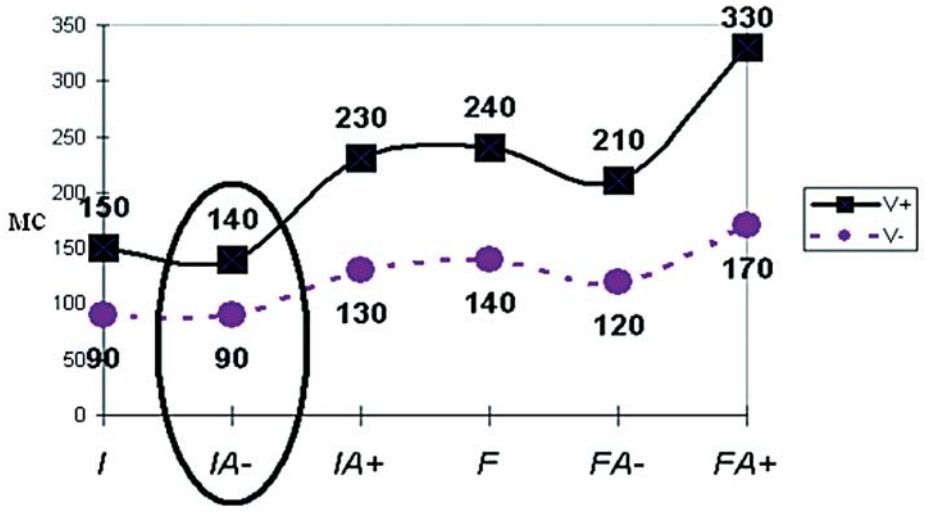


Рис. 3. Средняя длительность гласных в заданных фразовых позициях в контексте перед звонкими и глухими согласными, в миллисекундах («V+» — звонкий правый консонантный контекст, «V–» — глухой правый консонантный контекст)

Увеличение длительности гласных в заданных фразовых позициях относительно длительности гласных в нейтральной позиции «*IA–*» статистически достоверно как в контексте перед звонкими, так и в контексте перед глухими согласными. Исключение составляет начальная ударная позиция

[7] О большей длительности гласных перед звонкими согласными в английском языке см., например, вступительную статью в [13].

без АВ «*I*» (для *V*–), т.к. длительность гласных в сравниваемых позициях оказалась одинаковой (см. табл. 2, где статистически достоверные различия отмечены подчёркнутым курсивом).

Таблица 2

Тип сегментного контекста	*V+*	*V+*	*V+*	*V+*	*V+*	*V-*	*V-*	*V-*	*V-*	*V-*
Позиции в синтагме	*I*	*IA+*	*F*	*FA-*	*FA+*	*I*	*IA+*	*F*	*FA-*	*FA+*
***t*-критерий**	*2,63*	*10,06*	*11,65*	*9,72*	*11,03*	0,79	*9,68*	*7,58*	*3,15*	*7,59*
***p*<0,05**	*0,01*	*0,00*	*0,00*	*0,00*	*0,00*	0,43	*0,00*	*0,00*	*0,00*	0,00

t-критерий — критерий значимости изменения длительности гласных,
p — вероятность появления случайной величины вне допустимого интервала варьирования значений в заданной выборке.

Относительное увеличение длительности гласных, в зависимости от позиции во фразе, изменяется в диапазоне от 7% до 136% в контексте перед звонкими согласными и от 0% до 89% — перед глухими (см. рис. 4). Как видно, в контексте перед звонкими согласными оно больше: в начальной акцентной позиции «*IA+*» — на 20%; в конечной позиции синтагматического ударения «*F*» — на 15%; в конечной безударной позиции «*FA-*» — на 17%; в конечной акцентной позиции «*FA+*» — на 47%.

Длительность гласных в начальной акцентной позиции «*IA+*» оказалась не только меньше длительности гласных в конечной акцентной позиции «*FA+*» (на 72% перед звонкими согласными и на 45% — перед глухими), но и меньше длительности гласных под нейтральным синтагматическим ударением «*F*» (на 7% в контексте перед звонкими согласными и на 14% — перед глухими). Это подчёркивает значимость конечной фразовой позиции для длительности отдельных элементов.

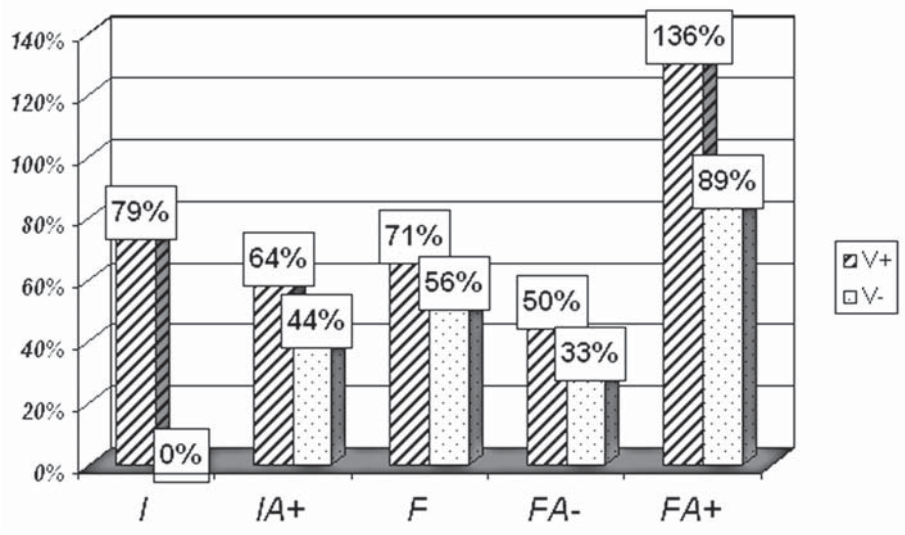


Рис. 4. Увеличение длительности гласных в заданных фразовых позициях в контексте перед звонкими и глухими согласными относительно длительности гласных в нейтральной позиции «IA–», в процентах

Максимальное увеличение длительности гласных в конечной акцентной позиции «*FA+*» можно объяснить действием «кумулятивного эффекта»[8]. Как было отмечено выше, по данным некоторых исследователей, увеличение длительности гласных под синтагматическим ударением достигает 30% [8]. Материал, исследованный в данной работе на участках гласных, показывает, что при АВ в конце фразы к известным 30% добавляется еще 100% в контексте перед звонкими согласными и 60% — перед глухими согласными.

Более детальный анализ статистически достоверных различий в длительности гласных показывает, что диапазон стандартного отклонения и доверительного интервала соотносятся по-разному для гласных в разных фразовых позициях и в разных сегментных контекстах.

Так, в контексте перед звонкими согласными (V+) в конечной акцентной позиции статистически значимыми различия являются за счёт полного расхождения величин стандартного отклонения, в результате чего вероятность совпадения длительности гласных в данном случае равна нулю (рис. 5, В). В отношении начальной акцентной позиции «*IA+*» следует отметить, что существует вероятность совпадения значений с показателями длительности гласных из нейтральной позиции, учитывая общую границу размахов стандартного отклонения в двух выборках (рис. 5, А). Таким образом, большее приращение длительности гласных соотносится с меньшей степенью вероятности совпадения значений.

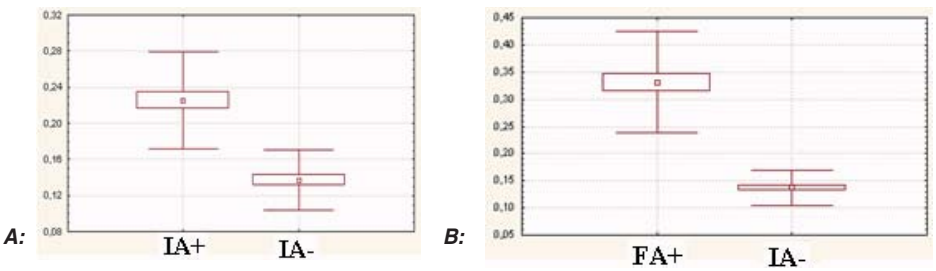


Рис. 5. Стандартное отклонение и доверительный интервал варьирования величин в заданных выборках

Разность в относительном приращении длительности гласных между акцентными позициями (72%) является статистически значимой, p=0,00 (рис. 6). Допустимые интервалы варьирования значений длительности в рассматриваемых позициях разные. Диапазоны стандартного отклонения при этом частично совпадают, что оставляет некоторую долю вероятности совпадения значений из двух выборок.

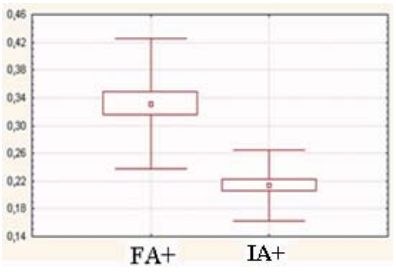


Рис. 6. Стандартное отклонение и доверительный интервал варьирования величин в заданных выборках

[8] О невозможности реализации «кумулятивного эффекта» см. в работе [14].

В контексте перед глухими согласными (V–) увеличение длительности гласных статистически достоверно в обеих акцентных позициях, благодаря разным значениям доверительных интервалов. Совпадение стандартных отклонений, однако, больше в начальной акцентной позиции «*IA+*» (рис. 7, A), чем в конечной «*FA+*» (рис. 7, B). В целом, совпадение стандартного отклонения по позициям больше, чем в контексте перед звонкими согласными (ср. рис. 7 с рис. 5), что можно объяснить меньшим приростом в относительной длительности гласных.

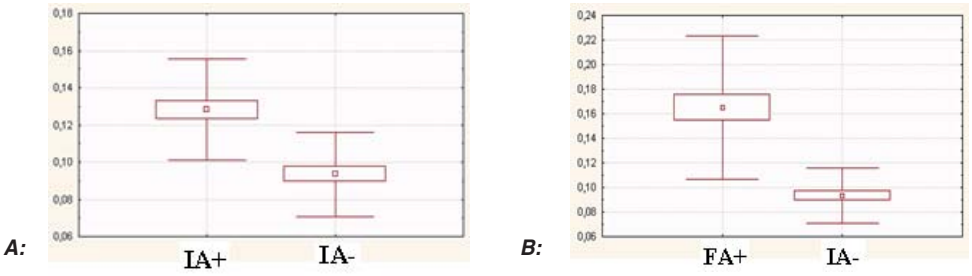


Рис. 7. Стандартное отклонение и доверительный интервал варьирования величин в заданных выборках

Разница в относительном увеличении длительности гласных между акцентными позициями «*IA+*» *vs*. «*FA+*» (45%) является статистически существенной (p=0,000214) (рис. 8), так же как и в контексте перед звонкими согласными. При этом области стандартного отклонения в акцентных позициях значительно совпадают (ср. рис. 8 с рис. 6). Статистическая существенность различий обеспечивается разными допустимыми интервалами варьирования абсолютных величин.

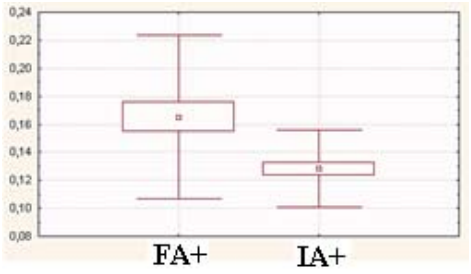


Рис. 8. Стандартное отклонение и доверительный интервал варьирования величин в заданных выборках

Рассмотрим сегментно-позиционные изменения длительности гласных в заданных фразовых позициях.

Разность в средней длительности гласных в контексте перед звонкими vs. глухими согласными варьируется от 50 до 160 мс (табл. 3). Наибольшая разность получена в позициях с АВ: «*IA+*» (100 мс) и «*FA+*» (160 мс) – и в позиции синтагматического ударения без АВ «*F*» (100 мс).

Таблица 3

Позиции в синтагме	***I***	***IA–***	***IA+***	***F***	***FA–***	***FA+***
***(V+)–(V–)*, мс**	60	50	100	100	90	160
***t*-критерий**	*10,36*	*6,87*	*10,10*	*11,93*	*12,99*	*9,94*
***p*<0,05**	*0,00*	*0,00*	*0,00*	*0,00*	*0,00*	0,00

Согласно t-критерию Стьюдента, полученная разница во всех случаях является статистически значимой. Диапазон стандартного отклонения показывает, что вероятность совпадения значений длительности гласных в сравниваемых выборках равна нулю (рис. 9). Следовательно, в акцентных позициях длительность гласных перед звонкими согласными всегда больше, чем перед глухими. Длительность гласных перед глухими согласными при появлении АВ не достигает значения длительности гласных перед звонкими согласными.

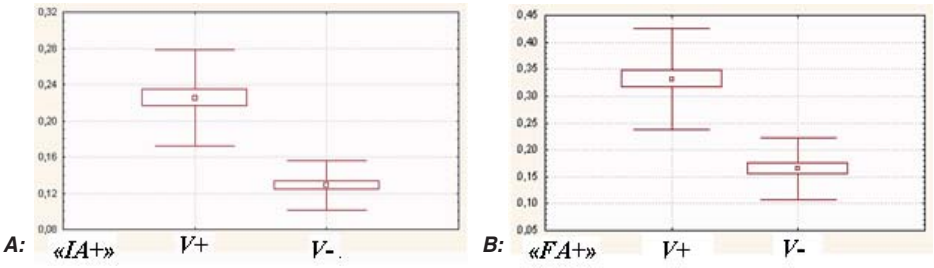


Рис. 9. Стандартное отклонение и доверительный интервал варьирования величин в заданных выборках

Относительное увеличение варьируется от 56% до 94%, в зависимости от позиции во фразе (рис. 10). При АВ разница в длительности гласных составляет 77% (в позиции «*IA+*») и 94% (в позиции «*FA+*»). Следовательно, длительность гласных в глухом сегментном контексте меньше длительности гласных в звонком сегментном контексте более чем на 50%[9].

Полученные данные позволяют предположить, что больший прирост абсолютной длительности гласных обусловливает большую разницу в сегментной длительности гласных (перед звонкими vs. глухими согласными). Так, например, относительное фразово-позиционное увеличение длительности гласных в конечной акцентной позиции («*FA+*») больше, чем в начальной позиции с АВ («*IA+*»). Разница в сегментной длительности гласных в конечной акцентной позиции также оказалась больше, чем в начальной акцентной позиции (94% vs. 77%) (см. рис. 10).

Выводы

Результаты исследования длительности гласных в условиях АВ позволяют сделать следующие выводы:

— относительное фразово-позиционное увеличение длительности гласных в АВ является значительным и статистически достоверным в контексте как перед звонкими, так и перед глухими согласными;

— в контексте перед звонкими согласными прирост длительности гласных больше, чем перед глухими, не только в абсолютных, но и в относительных показателях;

— в конечной акцентной позиции фразово-позиционное увеличение длительности гласных вдвое больше, чем в начальной, в контексте как перед

[9] Следует отметить, что сегментные различия в длительности гласных не выполняют функции различения гласных фонем в английском языке (ср.: «In most varieties of English, variations in lengths are completely allophonic» [9, с.225].

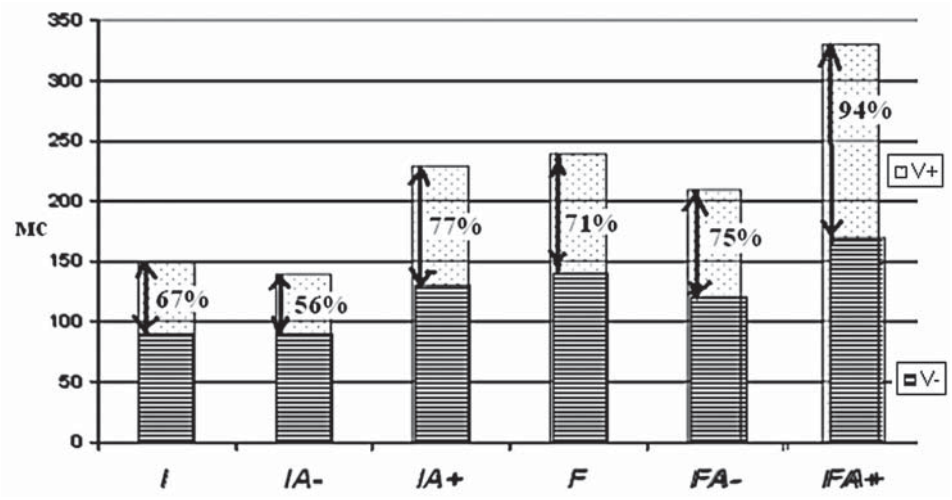


Рис.10. Увеличение длительности гласных в контексте перед звонкими согласными, по сравнению с длительностью гласных в контексте перед глухими согласными, в процентах

звонкими, так и перед глухими согласными; различия между начальной vs. конечной акцентными позициями статистически значимы;

— при появлении АВ в конечной фразовой позиции наблюдается действие «кумулятивного эффекта», когда к известным 30% [8] в позиции нейтрального синтагматического ударения может прибавиться ещё вплоть до 100% длительности гласных в контексте перед звонкими согласными и вплоть до 60% в контексте перед глухими согласными; таким образом, можно предположить, что именно эти дополнительные проценты ответственны за реализацию АВ в конечной фразовой позиции;

— в акцентных позициях сегментные различия по длительности становятся более яркими. Согласно полученным данным, в акцентных позициях длительность гласных перед звонкими согласными больше, чем перед глухими, на 77%–94%, а в неакцентных позициях — на 56–75%. Среди акцентных позиций более яркие сегментно-позиционные различия проявляются в конце фразы, где происходит наибольшее увеличение абсолютной длительности гласных как перед звонкими, так и перед глухими согласными.

Таким образом, при акцентном выделении вместе с увеличением длительности к концу фразы очень чётко проявляется аллофонное варьирование сегментно-позиционной длительности гласных. Полученные результаты свидетельствуют о совместном влиянии просодических и сегментных факторов на участках реализации АВ и вполне соответствуют выводам Л.А. Чистович и др.: «Реализация логических ударений производится путём таких преобразований артикуляторной программы произносимых предложений, которые не затрагивают её структурного существа, но могут быть описаны как введение некоторых коэффициентов, производящих закономерную деформацию элементов этой программы» [15].

Литература

1. *Скрелин П.А.* Фонетические аспекты речевых технологий. Дисс. докт. филол. наук. СПб.: Изд-во СПбГУ, 1999. 78 с. С. 3–5.
2. *Кодзасов С.В., Кривнова О.Ф.* Общая фонетика. М., 2001. 592 с. С. 515–524.

3. Потапова Р.К. Язык, речь, личность. М.: Языки славянской культуры, 2006. 491 с. С. 11.
4. Николаева Т.М. Семантика акцентного выделения. М., 2004.
5. Klatt D. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence // *JASA*, 59, 1976, pp.1208–1221.
6. Fry D.B. (1955). Duration and intensity as physical correlates of linguistic stress // JASA, 27, 765–768.
7. Гордина М.В. История фонетических исследований (от античности до возникновения фонологической теории). СПб., 2006. 538 с. С. 306–335.
8. Gimson A.C. An introduction to the pronunciation of English. London, 1966. 294 p. С. 266.
9. Ladefoged P. A course in phonetics. San Diego, 1982. 30 p.
10. Klatt D.H. Vowel lengthening is syntactically determined in a connected discourse // *Journal of Phonetics* 3, 1975, pp.129–140.
11. Halliday M.A.K. Intonation and Grammar in British English. The Hague, 1967. 62 с. С. 15.
12. http://ru.wikipedia.org/
13. Jones D. An English Pronouncing Dictionary, 17th edn., P.Roach, J.Hartman and J.Setter (eds.). Cambridge: CUP, 2006.
14. Cooper W.E., Eady S.J., Mueller P.R.J. Acoustical aspects of contrastive stress in question-answer context // JASA, (1985), Vol. 77, No. 6, pp. 2142–2156.
15. Чистович Л.А. и др. Речь: артикуляция и восприятие. М.; Л., 1965.

Филясова Юлия Анатольевна
аспирант кафедры фонетики и методики преподавания иностранных языков СПбГУ, младший научный сотрудник Лаборатории экспериментальной фонетики им. Л.В. Щербы Института филологических исследований (ЛЭФ ИФИ).
yuliyafill@gmail.com

Распределённая система фоноучёта «VoiceNet ID»

Тимофеев А.В.,
доктор технических наук

Введение

Опыт применения систем автоматизации фоноучёта и экспресс-исследований фонограмм речи серии «Трал» позволили специалистам ЦРТ сформулировать требования к распределённой системе фоноучёта «VoiceNet ID».

Основное назначение системы «VoiceNet ID» — хранение и оперативный поиск фонограмм, содержащих речь интересующего лица (или группы лиц).

Ключевым отличием новой системы от существующих комплексов является уникальная методология принятия обобщённого решения на основе результатов работы трёх независимых методов, а также оригинальная, трёхзвенная, архитектура программно-аппаратного комплекса (с тонким клиентом), по сути дела представляющего собой универсальное хранилище медиаданных.

Описание работы и функциональные возможности системы

Основными задачами, которые ставятся перед распределённой системой фоноучёта, являются:

- ведение федерального фоноскопического учёта для государственных органов, производящих расследование по уголовным делам;
- проведение фоноскопических исследований в интересах оперативно-разыскных подразделений;
- осуществление консультационной и учебно-методической деятельности.

В соответствии с требованиями распределённого фоноучёта система «VoiceNet ID» обеспечивает централизованное хранение и обработку информации. Ввод информации, формирование запросов на поиск (идентификацию) и проверку (верификацию) осуществляется с большого количества удалённых мест, при этом собственно доступ к данным осуществляется через Web-интерфейс. Таким образом, требования к оборудованию терминальных точек доступа к системе «VoiceNet ID» минимальны.

Система представляет собой распределённый вычислительный комплекс, включающий в себя вычислительное ядро и удалённые клиентские места. Архитектура системы обеспечивает функционирование клиентских мест по каналам Интернет с использованием Web-интерфейса. Система использует программный протокол, обеспечивающий возможность интеграции с другими информационными системами для

обмена информацией. Ядро системы функционирует под управлением операционной системы семейства UNIX. Безопасность передачи данных по открытым сетям обеспечивается протоколом HTTPS и авторизацией пользователей.

Состав системы и назначение её отдельных компонентов приводятся в таблице 1.

Таблица 1

Название компонента системы	Назначение
Брандмауэр	Обеспечение сетевой безопасности внутренней инфраструктуры вычислительного ядра
Web-сервер	Обеспечение инфраструктуры интерфейса пользователя
Диспетчер задач	Обеспечение диспетчерских функций системы
Хранилище данных	Обеспечение хранения данных и доступа к ним по запросу
Вычислительный кластер	Обеспечение параллельной отработки заданий пользователей
АРМ администратора	Автоматизированное рабочее место администратора
Фонолаборатория	Один из типов клиентов — криминалистическая лаборатория
Web-клиент	Наиболее массовый тип клиента — регистрация новых лиц в хранилище ядра системы, а также формирование запросов на поиск и проверку

Принцип работы системы основан на выделении из фонограмм речи и последующем попарном сравнении биометрических признаков (содержащихся в голосе индивидуальных, идентификационно значимых, признаков личности).

Выделение и сравнение индивидуальных признаков производится с использованием трёх различных языко- и текстонезависимых методов. В качестве основного используется спектрально-формантный метод, в качестве вспомогательных — метод статистик основного тона и метод на основе СГР (смеси Гауссовых распределений).

Необходимость использования одновременно трёх независимых методов обусловлена ограниченной областью применения каждого метода в отдельности, что иллюстрирует таблица 2 (количество знаков «+» отражает степень зависимости метода от параметров сигнала).

Таблица 2

Метод	Параметры сигнала		
	Продолжительность	**Качество сигнала**	**Физическое и эмоциональное состояние**
Спектрально-формантный	**+++**	**++**	**+**
СОТ	**++**	**+**	**++++**
СГР	**+++**	**++++**	++

Используемые методы поиска и параметры их надёжности

В системе «VoiceNet ID» используются языко- и текстонезависимые технологии поиска по голосу. Другими словами, неважно, что и на каком языке говорит человек. Исключение составляют т.н. тональные языки (вьетнамский, китайский, японский и т.п.), которые требуют перенастройки алгоритмов идентификации.

1. Спектрально-формантный метод

1.1. Данный метод основан на тезисе об уникальности геометрии речевого тракта у каждого человека и отражении данного факта в различных спектральных характеристиках речи. Наиболее явно различие спектральных характеристик проявляется в частотной ориентации и взаимном расположении формант.

1.2. Используемый в системе спектрально-формантный метод основан на выделении и сравнении положения и динамики поведения трёх и более формант. Данный метод защищён российским патентом.

1.3. Применение спектрально-формантного метода обеспечивает значение EER~8%. Значение EER для конкретного случая зависит от длительности и качества сравниваемых речевых фрагментов.

1.4. Данный метод является основным по следующим причинам:

1.4.1. Метод предъявляет самые низкие, по сравнению с другими, требования к качеству сигнала. Возможна работа с сигналами вплоть до отношения сигнал/шум 12 дБ.

1.4.2. Метод демонстрирует сравнительно высокую скорость выделения биометрических признаков и относительно робастен к типу канала.

2. Метод статистик основного тона

2.1. Данный метод использует шестнадцать различных характеристик основного тона (ОТ) голоса: среднее значение ОТ, максимальное значение, минимальное значение, медиана, процент участков с возрастающим тоном, дисперсия логарифма тона, асимметрия логарифма тона, эксцесс логарифма тона и другие параметры.

2.2. Значение EER для метода статистик основного тона зависит от длительности сравниваемых речевых фрагментов и может достигать величины ~16%. Необходимо отметить, что реализация данного алгоритма стала возможной благодаря созданию специалистами ЦРТ полностью автоматического высокоточного выделителя основного тона.

2.3. Достоинством данного метода является высокая скорость сравнения признаков и, как следствие, высокая скорость поиска или проверки личности. В то же время зависимость надёжности данного метода от эмоционально-психологического состояния диктора в момент произнесения позволяют использовать его в «VoiceNet ID» лишь в качестве вспомогательного.

3. Метод на основе СГР

3.1. Модели Гауссовых смесей (Gaussian Mixture Models) или смеси Гауссовых распределений на данный момент являются наиболее распространённым подходом к решению задач текстонезависимой идентификации.

3.2. Суть метода состоит в моделировании дикторозависимых акустических особенностей в пределах индивидуальных фонетических звуков (классов), которые входят в состав речевого сигнала. Сравнивая дикторозависимые акустические особенности в произ-

несении одного диктора с акустическими особенностями произнесения другого диктора, можно получить меру отличия дикторов в пространстве признаков.

3.3. Значения EER для метода на основе СГР зависят от длительности сравниваемых речевых фрагментов и могут достигать величины ~4–5%.

3.4. Высокая требовательность метода к качеству сигнала, высокая зависимость от обучающего материала, а также относительно большие временные затраты на выделение биометрических признаков не позволяют использовать его в системе в качестве основного. Впрочем, использование связки SVM-GMM позволило решить задачу каналокомпенсации путём понижения показателя EER на 2–3% по сравнению со связкой NN-GMM, используемой традиционно.

Аппаратные характеристики

Для обеспечения максимальной производительности системы целесообразно использовать блейд-серверы IBM на 19-дюймовом шасси. Сервер хранения базы данных рассчитывается с учётом объёма, занимаемого учётными карточками, и требований обеспечения быстродействия при обработке запросов. На хранение одной учётной карточки в базе данных отводится 7,5 МБ, из которых:

- 5,5 МБ — на хранение звукового сигнала, средняя продолжительность которого 4 мин. (включая паузы; в формате ИКМ 16 бит при частоте дискретизации 11025 Гц);
- 500 кБ — на хранение биометрической информации;
- 500 кБ — на хранение дополнительной информации о личности (установочные данные, информация из других информационных систем);
- 1 МБ — резерв.

Для реализации в полном объёме требуемых характеристик системы (что на практике означает проведение свыше 700 млн. сравнений в сутки) может потребоваться до 60 восьмиядерных блейд-серверов и сервер базы данных на 15 ТБ. Энергопотребление комплекса (с резервом) составит 100 кВт. Все серверы оснащаются блоками бесперебойного питания. С учётом постоянного повышения производительности компьютеров и совершенствования алгоритмов идентификации заявленное количество серверов может корректироваться.

В высокой степени надёжность, безопасность, высокая производительность комплекса обеспечиваются применением СУБД Oracle. В частности, СУБД Oracle позволяет работать с информацией практически неограниченному числу пользователей (при наличии достаточных аппаратных ресурсов), не проявляя тенденции к снижению производительности системы при резком увеличении их числа.

Механизмы масштабирования СУБД Oracle последней версии позволяют практически безгранично увеличивать мощность и скорость работы сервера базы данных и приложений простым добавлением новых узлов (серверов) кластера. Это не требует остановки и модернизации уже работающих приложений. Кроме того, выход из строя отдельных узлов кластера также не приводит к остановке приложения.

Для защиты кластера от перебоев электропитания предусмотрено его оснащение системой непрерывного питания.

Суточная загрузка системы представлена в таблице 3.

Таблица 3

Суточная нагрузка «VoiceNet ID»

Число вычислительных «лезвий»	8	18	38	58
Число ядер на «лезвии»	16	16	16	16
Общее число ядер	128	288	608	928
Максимальная расчётная загрузка кластера в сутки (количество попарных сравнений)	27 648 000	62 208 000	160 000 000	600 000 000

Заключение

Разработанные ЦРТ передовые средства и методы обработки речевых сигналов позволяют уже сегодня автоматизировать, с использованием распределённой системы фоноучёта, большую часть операций, связанных с поиском диктора и предварительным идентификационным исследованием по голосу и речи.

Тимофеев А.В.
Доктор технических наук, ООО «Центр речевых технологий», г. С.-Петербург.

Новые возможности анализа сигнала для определения положения формант в АПК «САПФИР»

Лобанова М.А.

В настоящее время АПК «САПФИР» активно используется для проведения фоноскопических экспертиз в лабораториях МВД. В то же время постоянно ведутся работы по его модернизации, включая не только дальнейшую автоматизацию работы эксперта, но и разработку новых методов проведения различных видов анализа сигнала. В данной статье будет дан краткий обзор некоторых новых возможностей программы, предлагаемых эксперту для определения положения формант исследуемого речевого сигнала.

Введение

В настоящее время при тесном сотрудничестве с ЭКЦ МВД России нами продолжается дальнейшее развитие созданного в 2007 году аппаратно-программного комплекса (АПК) для автоматизации проведения фоноскопической экспертизы «САПФИР» ([1]).

Сделаем краткий экскурс в историю создания данного комплекса.

К разрабатываемому АПК «САПФИР» предъявлялись следующие требования:

- возможность выполнения основных этапов типового технологического процесса производства экспертизы;
- возможность использования результатов уже выполненных работ на каждом последующем этапе выполнения экспертизы;
- автоматизация работы эксперта с целью сокращения временных трудозатрат и облегчения выполнения различных видов работ.

При разработке АПК «САПФИР» учитывалась принятая в настоящее время в МВД России методика проведения фоноскопической экспертизы, а именно методика «Диалект». Соответственно данной методике была разработана модульная структура программного обеспечения АПК «САПФИР», позволяющая не только удобным образом следовать методике «Диалект» (в том

числе обеспечить её освоение начинающими экспертами), но и разделить работу по проведению экспертизы между различными экспертами. Также в результате консультаций и плодотворного сотрудничества с экспертами ЭКЦ были определены наиболее важные направления автоматизации процесса производства экспертизы, что и было реализовано в конечном программном продукте.

В настоящее время АПК «САПФИР» активно используется для проведения фоноскопических экспертиз в лабораториях МВД. В то же время нами постоянно ведутся работы по его модернизации, включая не только дальнейшую автоматизацию работы эксперта, но и разработку новых методов проведения различных видов анализа сигнала. В статье дан краткий обзор некоторых новых возможностей программы, предлагаемых эксперту для определения положения формант исследуемого речевого сигнала.

Задача оценивания положения формант

Следует сразу отметить, что описываемые ниже способы определения положения формант не являются автоматическими и требуют активного участия эксперта.

Задача оценивания положения формант встаёт перед экспертом при проведении сравнительного лингвистического анализа образцов речи, а также при проведении сравнительного акустического микроанализа. При проведении микроанализа, заключающегося в поиске в образцах речи и в спорных фонограммах сопоставимых триад звуков (согласный–гласный–согласный) и их последующем параметрическом описании, требуется особенно точная оценка частотного положения формант.

Для оценивания значений формант эксперт, пользуясь возможностями имеющегося в его распоряжении ПО, обычно проводит вычисление спектрограммы с последующим анализом её изображения. Среди основных задач, которые при этом решает эксперт, можно назвать следующие:

- выбор оптимального размера спектрального окна для вычисления спектрограммы для исследуемого речевого фрагмента (с учётом значения основного тона голоса);
- выделение формантных траекторий на изображении вычисленной спектрограммы;
- оценка значений положений формант для выбранного в речевом сигнале момента времени.

Возможность и сложность выделения формантных траекторий по изображению спектрограммы, конечно, во многом зависят от характеристик исследуемого речевого сигнала: от качества сигнала (уровня шума и его амплитудно-частотных характеристик), от речевых навыков и манеры диктора, от экстралингвистических факторов, оказавших влияние на речевой процесс. Вместе с тем задачу выделения формантных траекторий можно облегчить программными средствами, среди которых, конечно, следует назвать возможность выбора для каждого конкретного сигнала оптимального диапазона цветопередачи спектрограммы, а также возможность вводить подъём амплитуды высокочастотных составляющих сигнала при их отображении.

В АПК «САПФИР», кроме перечисленных выше способов улучшить изображение спектрограммы и проявить на ней формантные полосы, реализованы дополнительные методы, которые условно можно разбить на три группы:

- методы, позволяющие одновременно (синхронно по частоте и по времени) проводить измерения в окнах, представляющих различные данные (результаты разных видов анализа);

- методы, основанные на сравнительном анализе спектрограмм идентичных речевых фрагментов;
- методы, основанные на различных способах отображения спектрограмм: например, дифференциальные спектрограммы или отображение спектрограммы с помощью «спектральных профилей» (см. рис.1).

Опишем кратко каждую группу методов.

Методы, позволяющие одновременно (синхронно по частоте и по времени) проводить измерения в окнах, представляющих различные данные (результаты разных видов анализа)

Данная группа методов базируется на возможностях пользовательского интерфейса АПК «САПФИР», позволяющего измерять синхронно в разных окнах значения частоты и времени.

На рис. 1 представлены виды окон осциллограммы, спектрограммы и мгновенного спектра сигнала, в которых отображены связанные между собой данные.

Для вычисления широкополосного спектра выбраны следующие параметры: размер спектрального кадра — 128 отсчётов сигнала, смещение между кадрами — 2 отсчёта.

В окне спектра отображены два мгновенных спектра, вычисленные для временных участков сигнала, имеющих общее начало, но разную длину (512 и 128 отсчётов сигнала). Начало этих временных участков соответствует положению поставленного в окне спектрограммы (осциллограммы) маркера.

Совершаемые экспертом, например, по окну спектрограммы перемещения указателя мышки отображаются также в окне осциллограммы и в окне спектра. В окне осциллограммы выделяются границы временного окна, для которого был вычислен спектральный срез, соответствующий указателю мышки в окне спектрограммы. В окне спектра отображается маркер, соответствующий частоте указателя мышки окна спектрограммы.

Возможность синхронного измерения временных и частотных координат данных, представленных в различных формах, позволяет эксперту принять решение о наличии форманты в той или иной частотной области и оценить её значение.

На рисунке показано проведение синхронного по частоте и времени анализа данных, представленных в разных окнах.

Методы, основанные на сравнительном анализе спектрограмм идентичных речевых фрагментов

Принятие решения по спектрограмме сигнала о том, где проходят формантные траектории, осложняется иногда не только плохим качеством фонограммы, но и ненормативными (в плане положения формант) особенностями речи.

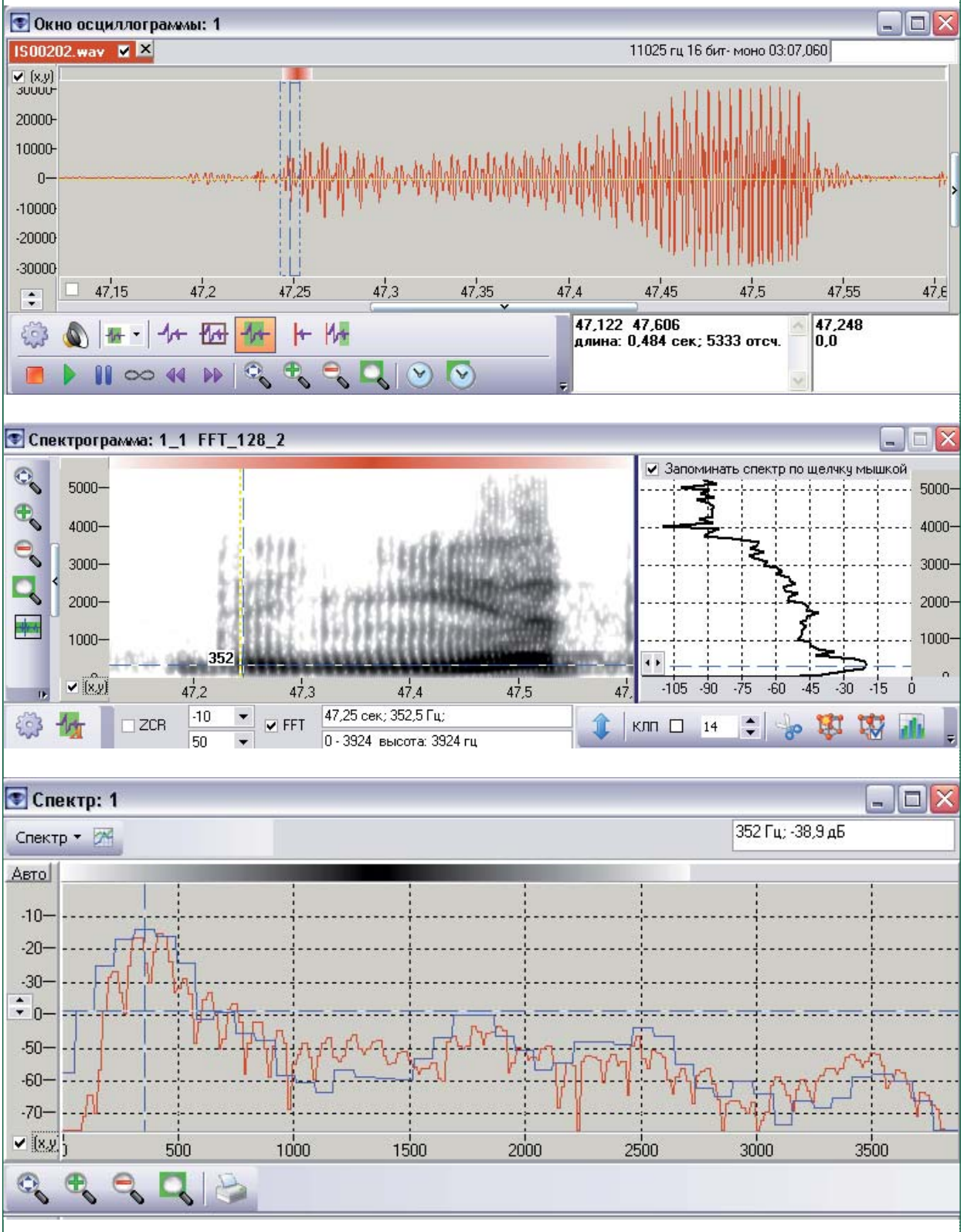


Рис.1. Окно осциллограммы с участком звукового сигнала; окно спектрограммы для данного сигнала; окно спектра с изображениями двух мгновенных спектров (широкополосного и узкополосного), вычисленных для временных кадров, имеющих общее начало

Мы имеем в виду наличие на спектрограмме сигнала выраженных затемнённых полос в тех частотных диапазонах, в которых их быть не должно согласно известным акустическим характеристикам (для присутствующих в речи фонетических единиц).

В этом случае провести формантный анализ может помочь сравнительный анализ спектрограмм идентичных речевых фрагментов (триад звуков) известного и спорного дикторов.

Одним из направлений развития АПК «САПФИР» явилась разработка возможностей проведения такого рода сравнений. На рис. 2 представлены наложенные друг на друга фрагменты спектрограмм идентичных речевых фрагментов, произнесённых разными дикторами. Пользовательский интерфейс позволяет эксперту сдвигать накладываемые фрагменты относительно друг друга по времени и по частоте, а также настраивать изображения, изменяя параметры цветопередачи для каждого из фрагментов отдельно.

На рис. 3 представлены наложенные друг на друга фрагменты кепстрограмм. Наложение кепстрограмм может помочь эксперту принять решение о возможности сравнения речевых сигналов (например, для определения идентичности эмоционального состояния сравниваемых дикторов, речь которых представлена на фонограммах).

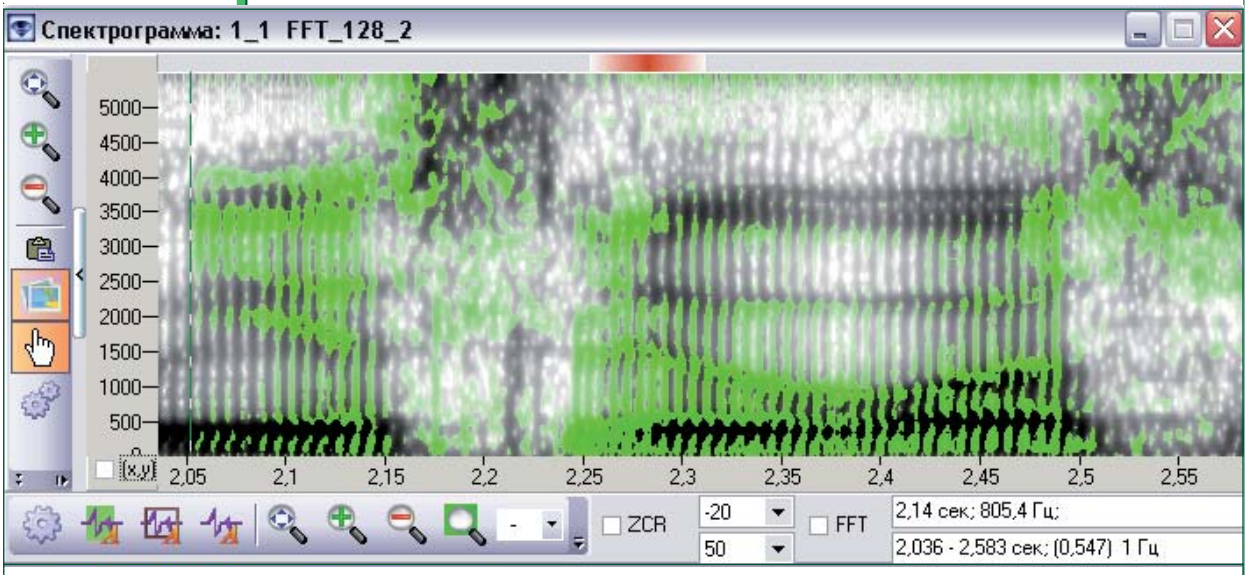


Рис. 2. Наложение друг на друга спектрограмм идентичных речевых фрагментов в АПК «САПФИР»

Методы, основанные на различных способах отображения спектрограмм (например, дифференциальные спектрограммы или отображение спектрограммы с помощью «спектральных профилей»)

Большую помощь эксперту при анализе сигнала, как показала практика, могут оказать дифференциальные спектрограммы, включённые в АПК «САПФИР» как один из возможных видов представления спектрограммы сигнала (см. [2]). Дифференциальные спектрограммы оказываются очень

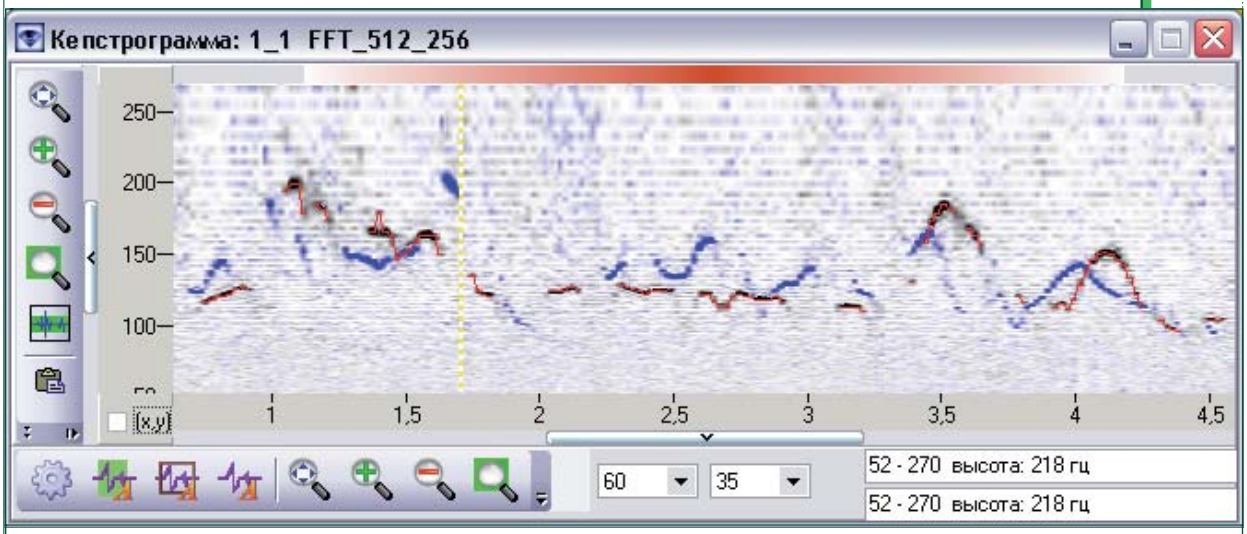


Рис. 3 . Наложение друг на друга кепстрограмм одинаковых фраз, произнесённых разными дикторами

полезными при определении границ между звуковыми единицами потока речи, а также для выделения на спектрограмме формантных траекторий.

Одним из направлений развития АПК «САПФИР» является реализация возможности отображения данных, представляемых на спектрограмме (вклада частотных составляющих в общий спектр сигнала, а также изменения этого вклада во времени), в виде «спектральных профилей».

Идея состоит в том, чтобы представлять изменение значения каждой частотной составляющей в виде отдельной линии на общем для всех частотных составляющих графике. На рис. 4 представлен такой график для нескольких соседних частотных составляющих.

Приведённые на рис. 4–6 изображения показывают возможности использования «спектральных профилей»:

- для определения положения формант (путём изучения графика трёх-четырёх соседних частотных компонент, соответствующих максимуму в спектре выбранного момента времени);
- для определения среднего и мгновенного значения основного тона голоса (по присутствующей в спектральных профилях амплитудной модуляции с частотой основного тона);
- для определения границ между звуковыми единицами речевого потока.

На графике «спектральных профилей» хорошо видно изменение энергии частотных характеристик, соответствующее происходящим в голосовом тракте говорящего перестройкам. Также на «спектральных профилях» хорошо видна периодическая структура, соответствующая работе голосовых связок.

На приведённых «спектральных профилях» в середине речи виден всплеск энергии частотных характеристик, соответствующий произнесению звука «р». На «спектральных профилях» хорошо видна разница в скорости нарастания (более пологая) и скорости уменьшения (более крутая) спектральной энергии частотных характеристик звука «р».

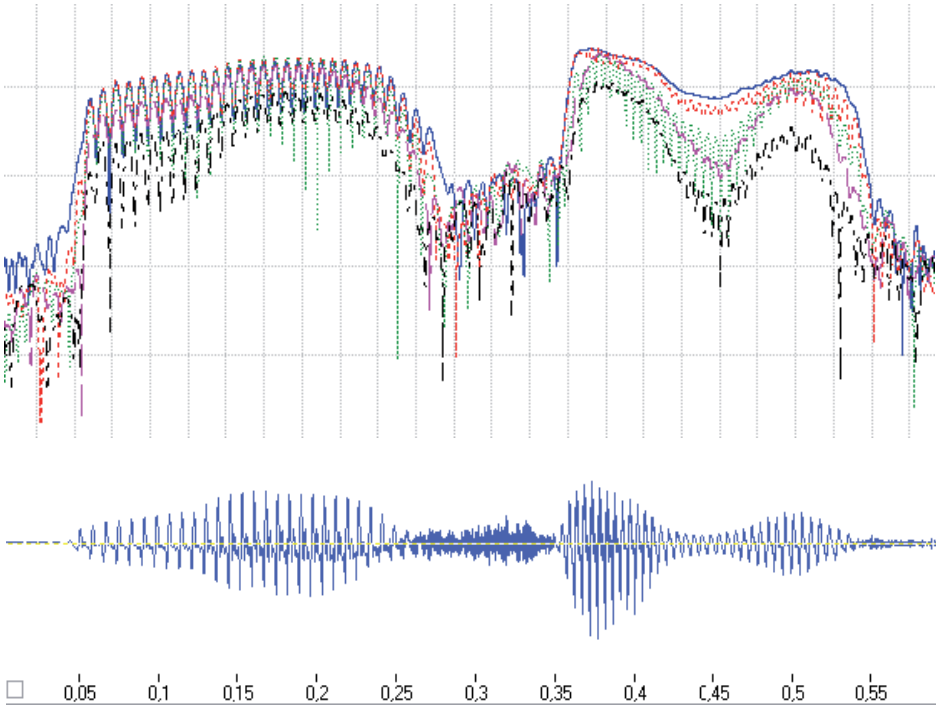


Рис. 4. График «спектральных профилей» соседних частотных компонент (верхнее окно): по горизонтальной оси отложено время, по вертикальной — значение спектральной энергии; в нижнем окне — осциллограмма того же сигнала

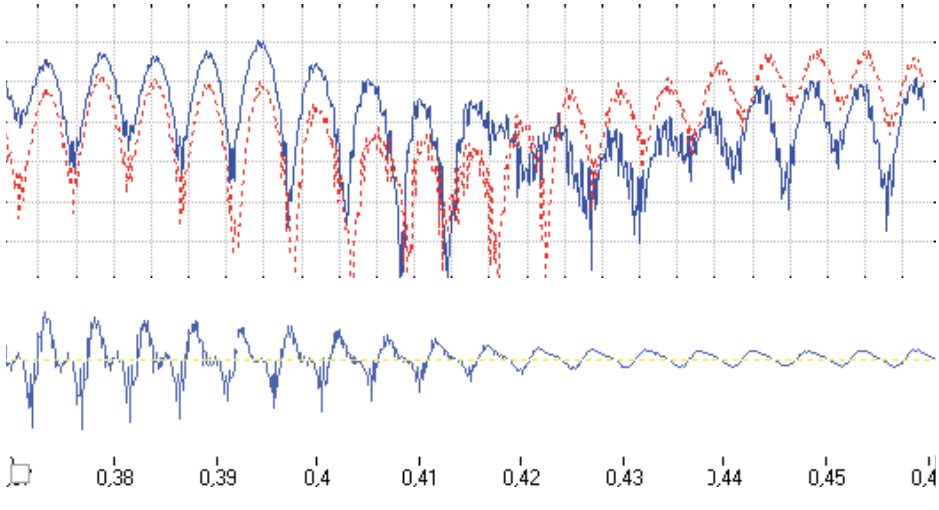


Рис. 5 . График «спектральных профилей» двух частотных компонент для небольшого временного окна (верхнее окно); в нижнем окне — осциллограмма того же сигнала

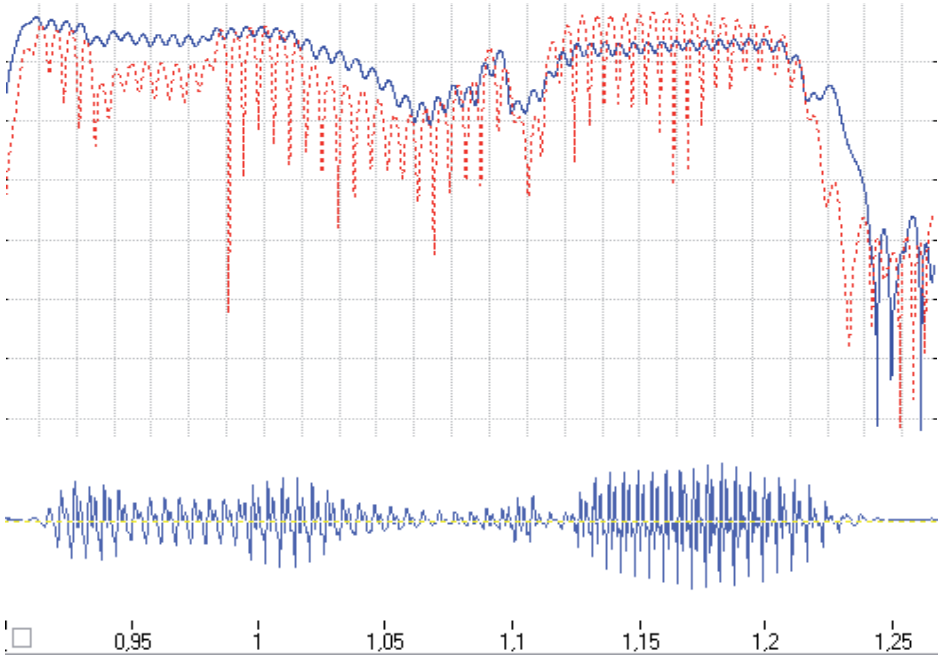


Рис.6. Осциллограмма сигнала (внизу) и «спектральные профили» двух частотных характеристик спектрограммы (наверху)

Заключение

Применение «спектральных профилей» для исследования отклика спектрального анализатора на речевой сигнал можно сравнить с применением микроскопа. На «спектральных профилях» хорошо видны изменения в речевом сигнале не только в энергетически выраженных частотных областях, но и во всех остальных. Как нам кажется, на «спектральных профилях» могут быть хорошо видны артикуляционные жесты говорящего, т.е. его стремление к перестройке артикуляционного тракта согласно речевой программе.

Возможно, изображение спектрограммы в виде «спектральных профилей» окажется интересным и полезным для проведения различных исследований речевого сигнала.

Литература

1. *Лобанова М.А., Назарова Т.В.* Универсальный комплекс для автоматизации проведения фоноскопической экспертизы (комплекс «САПФИР»). Материалы 15-й Международной научной конференции «Информатизация и информационная безопасность правоохранительных органов», 23–24 мая 2006 г. 313 с.

2. *Лобанова М.А.* Решение проблемы поиска идентичных речевых фрагментов в универсальном комплексе для автоматизации проведения фоноскопической экспертизы «САПФИР». Построение дифференциальных спектрограмм. Материалы 17-й Международной научной конференции «Информатизация и информационная безопасность правоохранительных органов», 20–21 мая 2008 г. 412 с.

Лобанова М.А.
ЗАО НПП «ИСТА-СИСТЕМС», Санкт-Петербург,
mal@ista.ru.

Компьютерный анализ звуковысотной системы голоса

Харуто А.В.,
кандидат технических наук

Интонационная составляющая речи, физическим носителем которой является мгновенная частота основного тона (ЧОТ), давно привлекает внимание исследователей как существенная психофизиологическая характеристика (см., например, [Lieberman, 1961; Женило, 1988, 1995]) и как филологический феномен [Кантер, 1988]. Анализ мелодического рисунка вокальной речи позволяет выделять «типовые» фрагменты исполнения — тоны, глиссандо, вибрато — и исследовать их характеристики [Харуто, 1998, 2005; Харуто, Смирнов, 1999; Смирнов, Харуто, 2000].

Анализ звукоряда на основе фонограммы предполагает проведение звуковысотной расшифровки, т. е. построения мелограммы (аналог контура ЧОТ; для удобства музыковедов мелограмма отображает ЧОТ в координатах высоты звука, а не частоты), с последующим её исследованием. Под звукорядом понимается набор звуков определённой высоты, на основе которых построена соответствующая музыкальная система. При исследовании предполагается, что наличие звукоряда проявляется в «более длительном» пребывании ЧОТ звука на определённых этим звукорядом уровнях, в то время как другие значения ЧОТ появляются в фонограмме только кратковременно — при переходе между частотами, относящимися к звукоряду. Выявление частот звукоряда возможно на основе одномерной плотности распределения ЧОТ: в соответствии с принципом максимального правдоподобия положения вершин локальных максимумов, распределения должны совпадать с частотами, образующими звукоряд.

Один пример мелограммы такого рода показан на рис. 1. В программе анализа музыкального звука SPAX, разработанной автором[1], предусмотрен режим отображения «сетки» звуковысотных ступеней при произвольном выборе их числа в октаве, а также возможность подстройки всей «сетки» по высоте; в данном случае наилучшим получилось совпадение высот, на которых «останавливается» голос, примерно с 19-ступенным равномерно темперированным звукорядом (т. е. содержащим 19 эквидистантных ступеней высоты в октаве).

Визуальный анализ характера контура ЧОТ достаточен для предварительных оценок и позволяет понять временную и звуковысотную структуру исследуемого процесса, однако потребность в более объективных данных заставляет разрабатывать алгоритмы анализа, дающие числовую оценку характеристик процесса.

[1] Программа SPAX. Свидетельство ФГУ «Роспатент» о регистрации № 2005612875 от 7 ноября 2005 г.

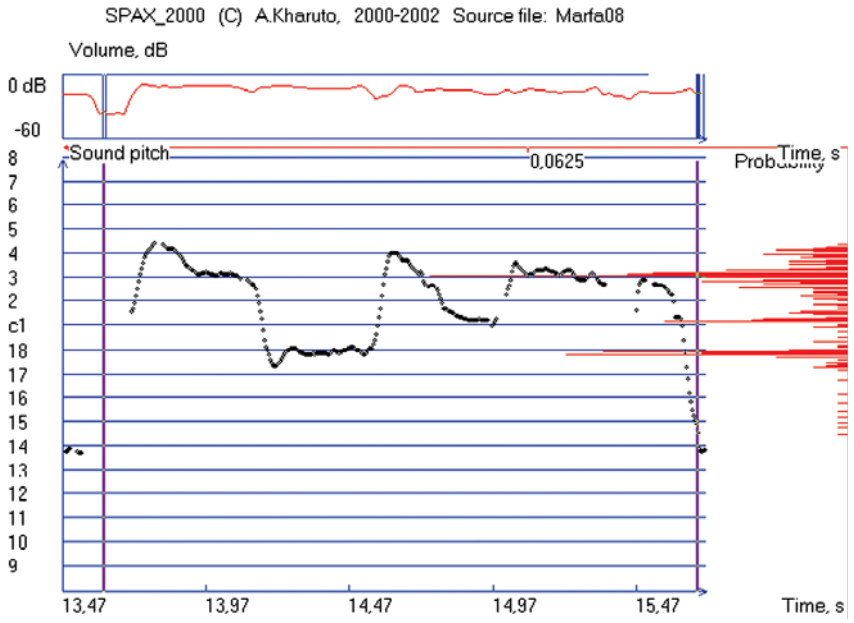


Рис. 1. Мелограмма фольклорного исполнителя, использующего равномерно-темперированный звукоряд примерно с 19-ю ступенями в октаве; диаграмма справа показывает результат анализа распределения «времени пребывания» звука на разных высотах

В примере на рис. 1 показана плотность распределения высоты звука для фрагмента фонограммы, выделенного вертикальными маркерами (график справа; ось вероятности направлена справа налево). Распределение имеет явно выраженные максимумы на тех «привычных» высотах голоса, где наблюдаются длительные горизонтальные участки в мелограмме.

Исследования распределения величины ЧОТ в речевых образцах показали, что оно часто оказывается полимодальным; в работе [Женило, 1988] отмечалось, что у некоторых дикторов моды распределения образуют систему, совпадающую по структуре с равномерно-темперированным музыкальным строем. Как показал ряд исследований автора доклада (см., напр., [Харуто, Смирнов, 1999; Смирнов, Харуто, 2000]), народные фольклорные певцы[2] обычно используют свой индивидуальный звукоряд с интервалом между звуками, меньшим, чем 1/12 октавы (практически от 1/17 до 1/30 и менее).

Следует отметить, что более полные данные могли бы быть получены путём исследования многомерных распределений, учитывающих статистические связи между высотами соседних звуков. Такие зависимости прослеживаются, например, в музыкальном исполнении на инструментах с нефиксированной настройкой: высота изменяется исполнителем по сравнению с «предписанной» нотами величиной для большего благозвучия (т. е. для исправления погрешностей 12-полутонового равномерно-темперированного строя). Измерения, подтверждающие это, были проведены разными исследователями и описаны, например, в работах [Рабинович, 1932; Сахалтуева, 1960; Parc, 1970].

Очевидно, что наличие звуковысотного вибрато и случайные или преднамеренные неточности выдерживания высоты «размывают» линию, соответствующую положению зву-

[2] Мы сознательно отличаем их от *профессиональных* исполнителей фольклора, которые часто имеют современное музыкальное образование и поют в 12-полутоновом равномерно-темперированном строе.

ковысотной ступени. При анализе фольклорных вокальных фонограмм, где вибрато отсутствует или появляется весьма редко (что можно проконтролировать путём просмотра всей звуковысотной расшифровки типа представленной на рис. 1), непосредственный анализ статистического распределения высоты звука будет давать необходимый результат; в случае более частого использования вибрато может быть произведено предусмотренное в программе SPAX интерактивное измерение каждого тона (т. е. звука с постоянной высотой) и тона, сопровождаемого вибрато. При этом фиксируется среднее значение тона на заданном интервале времени, а также параметры вибрато [Харуто, 2005], и дальнейшее статистическое исследование проводится по этим данным. Ниже мы ограничимся исследованием фонограмм, в которых отсутствует преднамеренное вибрато; будут рассмотрены методы и результаты анализа звуковысотной системы в фольклорном пении, близком к речитативу; для проверки и отладки алгоритмов использован образец фонограммы с музыкально-инструментальным исполнением.

В разработанной автором программе SPAX для определения ЧОТ применяется метод кепстра. Экспериментальная оценка точности определения ЧОТ по синтетическому сигналу показала отклонение от заданной частоты в пределах примерно в 4–5 центов (напомним, что октава составляет 1200 центов, а стандартный полутон равен 100 центам). Использование программы для звуковысотной расшифровки нескольких десятков образцов вокального фольклора не выявило никаких разночтений по сравнению со слуховым анализом фонограмм, проводившимся экспертами-фольклористами.

При исследовании распределения ЧОТ гистограмма строилась из «окон» размером $\Delta h = 5$ центов и (иногда) более. Известно, что размер окна гистограммы влияет на точность её оценивания. Чем меньше размер окна, тем выше точность оценки позиционирования элементов распределения (напр., требуемых в нашем случае локальных максимумов), т. е. меньше систематическая погрешность оценки с помощью гистограммы, использующей замену истинного распределения $\Psi(h)$ системой из N_H прямоугольных окон шириной Δh. Однако уменьшение размера окна приводит к меньшему числу зарегистрированных в нём значений процесса, т. е. меньшей вероятности p_i пребывания процесса в пределах этого i-го окна, что, в свою очередь, приводит к увеличению относительной среднеквадратичной погрешности оценивания значения $\psi(h_i)$ при данной высоте звука h_i, определяемой как (см., например, [Мирский, 1972, с. 313]):

$$\varepsilon^2 = \frac{1}{N} \times \frac{1 - p_i}{p_i},$$

где N — число некоррелированных выборок процесса.

Для выявления локальных максимумов распределения, соответствующих «привычным» частотам исследуемого голоса, могут быть использованы разные подходы. В частности, можно пытаться непосредственно зафиксировать локальные максимумы плотности распределения высоты звука (напомним, что высота пропорциональна логарифму ЧОТ). Для определения точек максимумов следует отыскивать в гистограмме $\{p_i, i = \overline{1, N_H}\}$ точки p_i, возвышающиеся над соседними, т. е. отвечающие условиям

$$p_{j-1} < p_j \quad \text{и} \quad p_j > p_{j+1}.$$

Поскольку положение максимума никак не привязано к границам окон гистограммы, для более точного определения его истинного положения целесообразно использовать аппроксимацию формы кривой $\Psi(h)$ в районе максимума. Например, в одном из исследованных нами алгоритмов через каждые три точки в районе максимума проводилась квадратичная парабола и далее аналитически определялось положение максимума. Очевидное ограничение на возможное расстояние между соседними обнаруживаемыми ступенями состоит в том, что этот интервал не может быть меньше $2 \times \Delta h$, что при $\Delta h = 5$ центам даёт величину минимального фиксируемого «шага» звуковысотной системы в 10 центов.

Такой алгоритм поиска обнаруживает, однако, все «выступающие» точки гистограммы, которых при анализе реального исполнения оказывается очень много и которые, по всей видимости, не являются ступенями звукоряда. Большая часть интервалов между «ступенями» при этом только ненамного превышает указанный нижний предел. Для примера рассмотрим анализ одной музыкальной фонограммы — это упражнение, «коряво» исполненное начинающим скрипачом (он проигрывал гаммы вверх и вниз). «Идеал», к которому стремился исполнявший, — ряд равноотстоящих по высоте «ступенек» на высотах, соответствующим нотам 12-полутонового равномерно-темперированного строя. Однако поскольку скрипка — инструмент с нефиксированной настройкой, здесь возможны (и реально присутствуют) погрешности интонирования. На рис. 2 показан фрагмент мелограммы и гистограмма распределения высот, оценённая по всей фонограмме. Здесь хорошо видна система «пиков» распределения, которые практически не перекрываются, но разнонаправлено сдвинуты по сравнению со стандартными высотами нот 12-полутонового звукоряда. «Пики» распределения имеют также разную ширину и иногда раздвоены, что объясняется нестабильностью высоты при исполнении — «дрожанием» в процессе исполнения одного звука (увеличенная ширина) и неточностью средней высоты при повторном проигрывании той же ступени (раздвоение).

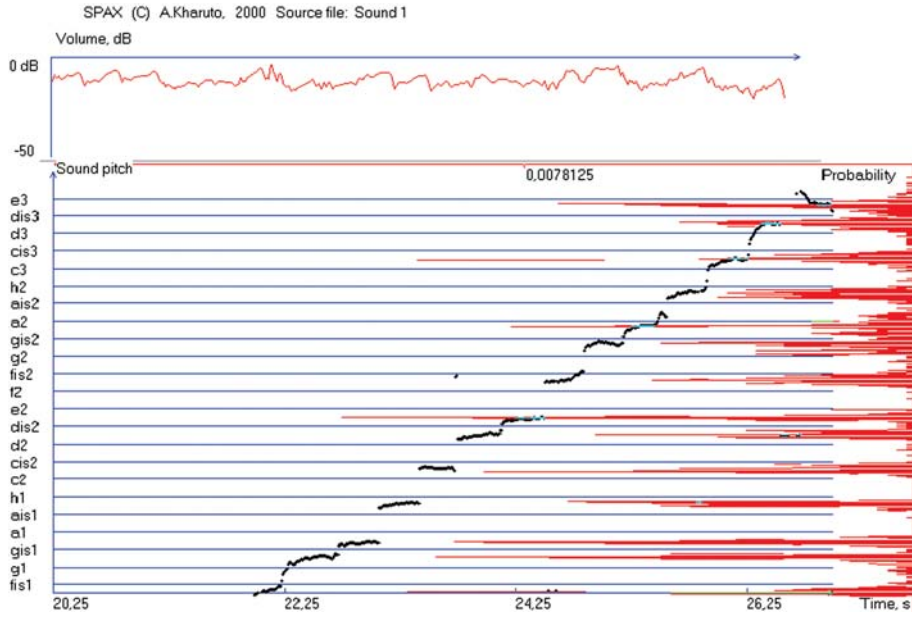


Рис. 2. Фрагмент звуковысотной расшифровки ученического исполнения на скрипке и оценка распределения высот

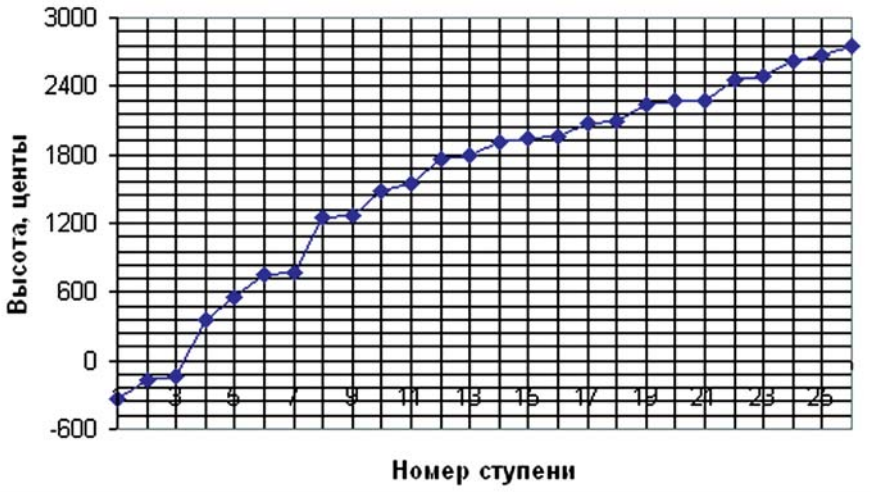


Рис. 3. Результат оценки звукоряда (см. пример на рис. 2) путём поиска всех локальных максимумов

Определение ступеней звукоряда путём фиксации всех локальных максимумов даёт результат, показанный на рис. 3 (за «нуль» высоты принята нота *до* первой октавы).

Здесь видны как «повторяющиеся» ступени, разделённые очень малыми интервалами (соответствующие, видимо, раздвоенным пикам распределения), так и переменные по величине «скачки» между ступенями. Присутствуют, соответственно, как очень мелкие шаги (соответствующие «сдвоенным» пикам), так и близкие к ожидаемым для данного случая, т. е. примерно кратные 100 центам.

Если об исследуемом звукоряде нет априорных сведений, то по подобным данным определить его структуру было бы затруднительно. Выделение «основных» ступеней по признаку наибольшей вероятности некорректно, поскольку суммарное время пребывания высоты звука в том или другом окне гистограммы, т. е. оценка вероятности «использования» каждой из искомых ступеней, существенно зависит и от исполняемой мелодии (а в случае речевого общения — от требуемой интонации высказывания), и от «качества» её воспроизведения, что иллюстрируется рис. 2. Таким образом, вполне возможно, что коротко прозвучавший тон окажется одним из основных, образующих звуковысотную систему, так что его нельзя не учитывать. Кроме того, в исполнении (фрагменте речи) могут отсутствовать некоторые ступени звукоряда, поскольку они «не нужны» в данном случае (но понадобятся в другом).

Используя предположение о том, что искомый звукоряд является равномерно-темперированным, т. е. что его ступени разнесены на равные интервалы по шкале высоты, можно предложить другой способ анализа, основанный на оценке всей гистограммы в целом и поиске в ней *периодически повторяющихся «пиков»*. Для такой оценки автором был (как и для определения ЧОТ) использован метод кепстра: гистограмма логарифмировалась (что

Рис. 4. Спектр, вычисленный для распределения высот звука (см. пример на рис. 2)

«уравнивает» в некоторой степени вклад в оценку часто и редко используемых ступеней), затем вычислялся её спектр. Для приведённого выше примера — ученического исполнения на скрипке — получается спектр распределения, показанный на рис. 4.

Наиболее мощные максимумы обнаруживаются в точках №№ 5, 7, 11 и 15; интенсивности соответствующих компонент спектра отображают «выраженность» данной периодической составляющей (т. е. совокупности «пиков», размещённых через соответствующий шаг по высоте). График интенсивностей для перечисленных наиболее выраженных периодических составляющих в зависимости от предполагаемых шагов между ступенями звукоряда показан на рис. 5. Здесь видно, что наиболее выраженной (наиболее вероятной) структурой в данном исполнении является звуковысотная система с шагом 100 центов (которая и «предписана» стандартными высотами нот 12-полутонового звукоряда). Ошибки исполнения порождают побочные пики, но их интенсивность намного меньше.

Рис. 5. Зависимость «выраженности» периодической компоненты звукоряда от шага между ступенями (см. пример на рис. 2)

Рис. 6. Спектр, вычисленный для распределения высот звука в фонограмме русского фольклорного пения (см. рис. 1)

Анализ первого из приведённых выше примеров (русское фольклорное пение в звуковысотной системе, содержащей, по предварительной оценке, примерно 19 ступеней в октаве, — см. рис. 1), даёт спектр плотности распределения, приведённый на рис. 6. На рис. 7 показаны значения интенсивности периодической компоненты распределения высоты для основных «пиков» данного спектра. Здесь число значительных и сопоставимых по величине максимумов значительно больше, однако для самого мощного из них (точка № 18) получается величина соответствующего шага по высоте, равная 58,8 центам, что соответствует 1200:58,8=20,4 ступеням в октаве. Другие пики спектра получаются из-за большого количества промежуточных по высоте звуков — в частности, из-за украшения мелодии движением вверх-вниз относительно средней высоты исполняемого звука.

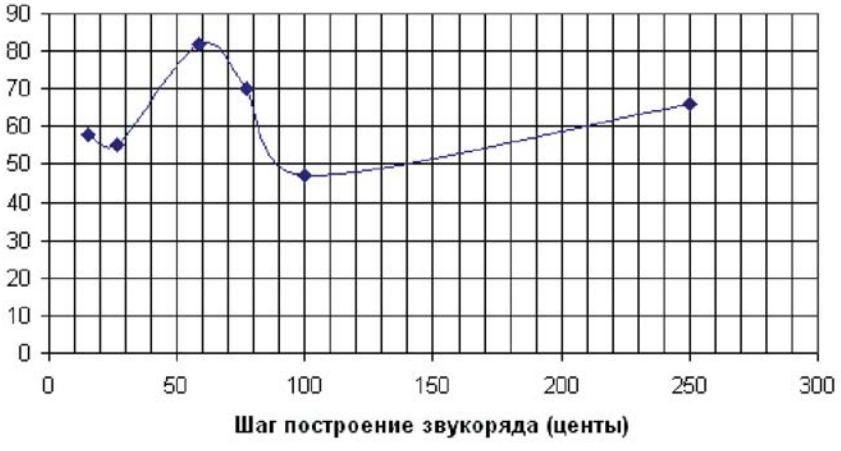


Рис. 7. Зависимость «выраженности» периодической компоненты звукоряда от шага между ступенями для образца русского фольклорного пения (см. рис. 1)

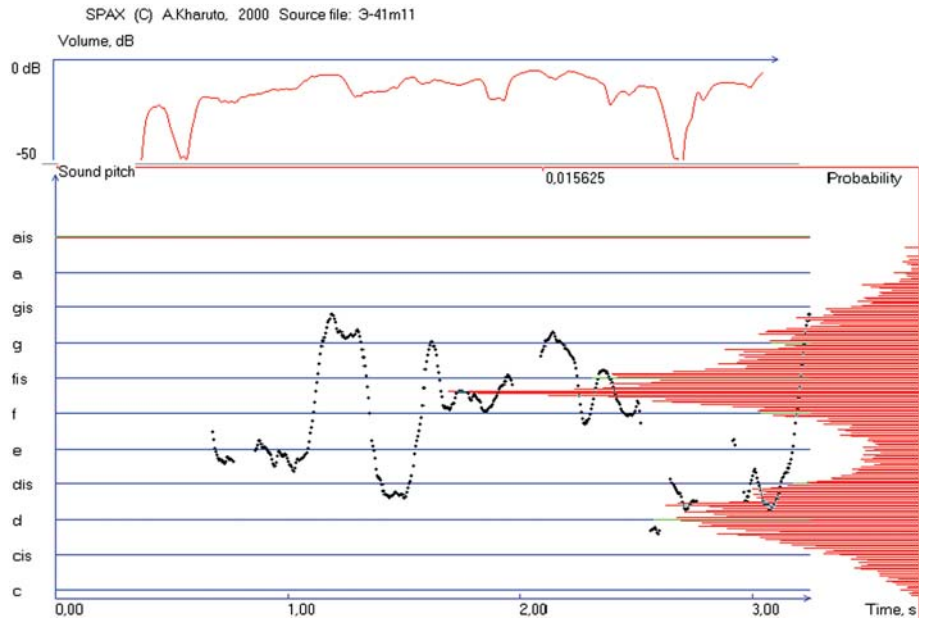


Рис. 8. Фрагмент звуковысотного рисунка и распределение высоты в эвенкийской песне

Ещё один пример анализа — эвенкийская песня (мужской голос, исполнение близко к речитативному). Звуковысотная расшифровка фрагмента и гистограмма распределения высоты показана на рис. 8. Спектр для этого распределения показан на рис. 9. Как и в предыдущем случае, спектр содержит много максимумов, что отображает сложную звуковысотную структуру[3]. На рис. 10 показаны интенсивности для наиболее выраженных периодических составляющих этого спектра.

При анализе спектра распределения высоты для этого образца выявляется основной по выраженности шаг звуковысотной системы, равный 20 центам (точка № 51 в спектре).

Таким образом, компьютерный анализ звуковысотной системы фонограммы на основе распределения высоты звука с последующим исследованием периодичности структуры этого распределения позволяет определить интервал, образующий равномерно-темперированный звукоряд исполнителя. По-видимому, можно также на основе характера спектра распределения (количества дополнительных «пиков», их интенсивности и пр.) интегрально оценивать точность следования звукоряду в исследуемом исполнении.

Отметим, что указанный тип звукоряда не является единственно возможным: так, в тувинском горловом пении (и сходных с ним монгольском, тибетском и др.), где во время вокализов слышны по меньшей мере одновременно два голоса на разных высотах, мелограмма «верхних» голосов показывает использование *натурального* звукоряда, где звуковысотные ступени разнесены на равные *по частоте* (а не по высоте) интерва-

[3] Здесь (как и на других графиках спектров) для удобства масштабирования «вырезана» постоянная составляющая, вследствие чего образовался искусственный максимум в точке № 2, — в расчётах он не учитывается.

Рис. 9. Спектр, вычисленный для распределения высот звука в фонограмме эвенкийского фольклорного пения (см. рис. 8)

лы. Это связано с используемым механизмом звукоизвлечения: все слышимые «голоса» образуются из обертонов нижнего основного звука, т. н. *бурдона*, высота которого во время исполнения практически неизменна (см, например, [Харуто, Карелина, 2008; Харуто, 2008].

Сопоставляя результаты нашего анализа с традициями европейского этномузыковедения, использующего 12-полутоновую нотацию (иногда — с дополнительными знаками микроальтерации, позволяющими фиксировать, например, четвертитоновые интервалы), можно заключить, что зарегистрирован-

Рис. 10. Зависимость «выраженности» периодической компоненты звукоряда от шага между ступенями для образца эвенкийского фольклорного пения (см. рис. 8)

ные в фольклорном пении интервалы между ступенями должны измеряться гораздо точнее и не могут быть отображены указанными средствами нотации.

Следует отметить, что представленные результаты носят предварительный характер и требуют дальнейшей проверки и сопоставления с данными, полученными экспертами-музыковедами с помощью «традиционных» слуховых методов.

Литература

1. *Lieberman Ph.* (1961) Perturbations in Vocal Pitch // The Journal of the Acoustic Soc. of America. 1961. v.33, N 5. — p. 597–603.
2. *Женило В. Р.* Анализ параметров частоты основного тона голоса человека для автоматической идентификации личности // Академия наук СССР, Вычислительный центр. Сообщения по программному обеспечению ЭВМ. М., 1988.
3. *Женило В. Р.* Компьютерная фоноскопия // М: Академия МВД России, 1995.
4. *Кантер Л. А.* Системный анализ речевой интонации: Учебн. пособие. М.: Высшая школа, 1988.
5. *Харуто А. В.* Компьютерный анализ звука в музыковедческом исследовании. Труды международного научного симпозиума «Информационный подход в эмпирической эстетике». Таганрог: Изд. ТРТУ, 1998.
6. *Харуто А. В., Смирнов Д. В.* Использование компьютерного анализа в исследовании звуковысотного строения народной музыки //Материалы международных конференций памяти А. В. Рудневой. М.: Московская гос. консерватория, 1999.
7. *Смирнов Д. В., Харуто А. В.* Нелинейный звукоряд в музыкальном фольклоре: общая закономерность и индивидуальность // Языки науки — языки искусства // Общ. ред. З.Е. Журавлевой, В. А. Копцик, Г. Ю. Резниченко. М.: МГУ, 2000.
8. *Харуто А. В.* Статистическое исследование характеристик вибрато // Сборник трудов XIV международной научной конференции «Информатизация и информационная безопасность правоохранительных органов». М.: Академия управления МВД России, 2005.
9. *Рабинович А. В.* Осциллографический метод анализа мелодии // Проблемы музыкознания. Теоретическая библиотека. М.: Музгиз, 1932.
10. *Сахалтуева О. Е.* О некоторых закономерностях интонирования в связи с формой, динамикой и ладом // Труды кафедры теории музыки Московской гос. консерватории им. П. И. Чайковского. Вып. 1. М.: Музгиз, 1960.
11. *Рагс Ю. Н.* О художественной норме чистой интонации при исполнении мелодии: Дисс. канд. искусствоведения. М., 1970.
12. *Мирский Г. Я.* (1972) Аппаратурное определение характеристик случайных процессов. Изд. 2-е перераб. и доп. М.: Энергия, 1972.
13. *Харуто А. В., Карелина Е. К.* К вопросу о музыкально-акустических свойствах тувинского горлового пения. «Музыкальная академия», 2008. № 4, С. 108–113.
14. *Харуто А. В.* Тувинское горловое пение: акустический анализ и модель звукообразования // Сб. трудов XX сессии Российского акустического общества, секция «Акустика речи» М., РАО, 2008. С. 106–110.

Харуто А.В.
Кандидат технических наук,
доцент Московской государственной консерватории.

Основные тенденции развития многоязычной корпусной лингвистики

(Часть 1)

Потапова Р.К.,
доктор филологических наук, профессор

Корпусная лингвистика (КЛ) — это многоаспектный раздел прикладной лингвистики, «обслуживающий» целый ряд отраслей теории и практики вербальной коммуникации на базе новых информационных технологий.

Термин «Корпусная лингвистика» (*Corpus Linguistics)* отражает характер объекта, с которым имеет дело данная эмпирическая область исследований. Сюда относятся, прежде всего, тексты на естественных языках в машиночитаемом формате, образующие массивы, коллекции, а также специально оформленные корпусы (*Corpora)* [Леонтьева, 2006].

Текстовые корпусы начали создавать ещё в 60–70 гг. XX века, т.е. корпусная лингвистика (КЛ) существует уже более 30 лет. За это время созданы десятки банков текстовых данных в первую очередь для английского, а затем и для других европейских языков и языковых пар; на основе текстовых корпусов (ТК) созданы сотни словарей (*corpus-based dictionaries).*

COBUILD — первый словарь, основанный на корпусных данных, – вышел в 1987 г. и был принят как стандарт с учётом требований теории и практики. С тех пор все современные словари, особенно предназначенные для изучения неродного языка, основываются на материале имеющихся и всё время пополняющихся ТК.

В 1995 г. вышел в свет Collins COBUILD English Dictionary (CCED), в котором отразились существенные изменения в языке относительно первых публикаций COBUILD: некоторые слова и значения выпали, но в то же время в него вошли американизмы и ряд технических терминов.

ТК создаются не только для основных европейских языков (английского, французского, немецкого), но также и для менее распространённых языков (шведского, норвежского, финского). В настоящее время существует значительное число ТК: разнотипных и разноразмерных, одно-, дву- и многоязычных, для письменного и устного вариантов языка. Создаются также параллельные корпусы: англо-норвежский, чешско-английский, словацко-

русский, словацко-хорватский и др. Более того, современные грамматики и словари формируются только на основе корпусной поддержки.

Согласно стандарту Британского национального корпуса (БНК) были созданы текстовые корпусы для многих европейских языков. Характеристика «национальный», служащая для конкретизации варианта языка, описываемого в корпусе, стала применяться для обозначения представительного ТК любого языка.

Как правило, национальный корпус — это отдельная комплексная система, образование и ведение которой требует больших трудозатрат как со стороны лингвистов, так и с учётом программного обеспечения. Современные ТК-системы выбирают и проводят определённую лингвистическую политику и используют для этого последние достижения информационных технологий.

Назовём только некоторые ТК (наиболее известные из них можно найти в Интернете): COBUILD (основан в 1980 г., руководитель Джон Синклер); British National Corpus (BNC), или БНК; Bank of English (Банк английского языка); ALEX (банк английской и американской литературы по западной философии); БК (Брауновский корпус); ICE (International Corpus of English); Longman/Lancaster Corpus; London-Lund Corpora; OED (Oxford English Dictionary); CUMBRE Corpus (корпус современного испанского); Чешский национальный корпус; Словацкий национальный корпус; Китайский текстовый корпус.

В России первым опытом создания большого лингвистического корпуса был Машинный фонд русского языка [Андрющенко, 1989], целью которого было создание представительного корпуса с подкорпусами различных жанров и соответствующих программных средств, а также комплексная информатизация лингвистических исследований, включая создание грамматик и словарей. Несмотря на то, что в полном виде программа выполнена не была, удалось собрать коллекции текстов разного типа, перевести в машинный вид многие традиционные словари.

В настоящее время Фонд обслуживает внутренние задачи Института русского языка (ИРЯ) РАН: ведение Русского диалектологического атласа, создание автоматического конкорданса для текстов русского фольклора, политических текстов, текстов древнерусских источников XI–XVII вв. и др. Каждая из задач требует создания отдельного пакета программ. В состав Машинного фонда входит большое количество словарей: «Грамматический словарь русского языка» А.А. Зализняка, «Русский орфографический словарь», «Русский синтаксический словарь» Е.А. Золотовой и др. В Фонд вошли также коллекции русской художественной литературы (М.Ю. Лермонтов, Ф.М. Достоевский), коллекции русских поэтических текстов. Руководитель Фонда А.Я. Шайкевич самой важной научной задачей считает проведение дистрибутивно-статистического анализа текстов и создание объективного описания языка, используя языково-независимый метод формирования «естественных» классов [Shaikevich, 1997; Рахилина, Шаров, 2003].

В начале XXI века в России начата работа по созданию представительных корпусов для русского языка. Два проекта — БОКР (Большой корпус русского языка) и РС (Русский стандарт), которые должны были представить русский литературный язык во всех значимых жанрах и видах использования [Шаров, 2003], слились с «Корпусом ЦЛД — МЕУ» [Сичинава, 2002]. Последний создаётся с 2001 г. общественной организацией ЦЛД (Центр лингвистической документации, руководитель – В.А. Плунгян). Была создана Ассоциация «Национальный корпус русского языка», в которую вошла большая группа лингвистов Москвы, Санкт-Петербурга, Новосибирска и других научных цен-

тров России. Планируемый объём корпуса — 200 млн. слов. Подробнее об Ассоциации, её участниках и планах можно посмотреть в Интернете на странице www.ruscorpora.ru.

Кроме того, отдельные коллективы РФ продолжают свои работы по созданию специальных ТК [Корпусная лингвистика, 2003]. В Санкт-Петербургском университете проводятся регулярные конференции, посвящённые КЛ.

Текстовые корпусы (ТК) могут формироваться по разным основаниям: авторские, по жанрам, стилям, по дате источника, по научным направлениям и т.п. Создатели ТК должны определять, какие порции и пласты языка нужно в них представить, что зависит от конкретных задач и внешних условий (например, финансирования), а также от адресатов ТК [Шаров, 2003; Рыков, 2003].

Что касается источников формирования текстовых корпусов (ТК), то в настоящее время проблем не существует благодаря Интернету, технологиям автоматического чтения и сканирования, быстродействию компьютеров, практически неограниченным объёмам памяти и т.д.

С 1996 г. стал выходить Международный журнал по корпусной лингвистике, на страницах которого обсуждаются разные аспекты формирования и ведения текстовых массивов, описываются новые ТК, обсуждаются вопросы их аннотирования. ТК, снабжённые лингвистической информацией, называют размеченными или аннотированными. Чем богаче разметка (например, морфологическая, синтаксическая), тем большую ценность имеет корпус.

Так, в частности, в Национальном корпусе русского языка используется пять типов разметки: метатекстовая, морфологическая, акцентная, синтаксическая и семантическая. Две последние выполняются на небольшом фрагменте корпуса.

Недавно стал создаваться аннотированный корпус для русских текстов в ИППИ РАН [Богуславский, Григорьев и др., 2000], который состоит из нескольких подкорпусов. Тексты последних различаются уровнем аннотации:

— лемматизированные тексты, в которых для каждого слова указывается его основная форма и часть речи;

— тексты с морфологической информацией, в которых для каждого слова указывается его основная форма, часть речи и полный набор морфологических характеристик;

— тексты с синтаксической информацией, в которых для каждого слова указывается его основная форма, часть речи и морфологические характеристики, а для каждого предложения — его синтаксическая структура.

Выполняемая автоматически разметка корректируется лингвистами.

К 2000 г. создано не меньше 20 аннотированных корпусов для основных европейских языков. Из них, по крайней мере, три — с синтаксической информацией. Наиболее известны Perm Treebank, созданный в Пенсильванском университете в 1990 г. [Markus, Santorini, Marcinkiewicz, 1993], и создаваемый по его образцу Пражский банк деревьев зависимостей (PDT – Prague

Dependency Treebank). Эти работы постоянно освещаются в Пражском бюллетене по математической лингвистике [Bohmova, 2001; Hajicova, Pajas, Vesela, 2002].

PDT — это исследовательский проект Карлова университета в Праге. Схема аннотирования включает три уровня: морфологический, аналитический и тектограмматический. На первом из них проводятся стандартные для всех систем операции лемматизации и определения всех морфологических характеристик (используется примерно 3000 значений морфологических тэгов) для словоформ входного текста. На втором уровне строится поверхностная синтаксическая структура, называемая analytic tree structure (ATS): это промежуточное дерево зависимостей, в котором каждое слово и знак препинания представлены отдельными узлами с приписанными им характеристиками теньеровского типа (субъект, объект, адвербиал, атрибут). Перевод из линейных структур (с их скобочной записью) в древесную проводится полуавтоматически. Такой метод был испытан и отработан на трансформации деревьев составляющих английского языка из Пенсильванского банка в деревья тектограмматического уровня, принятые в PDT. Третий уровень строит тектограмматическую древесную структуру (Tectogrammatical Tree Structure — TGTS), представляющую собой глубинное синтаксическое дерево предложения. В нём в качестве узлов остаются только полнозначные слова; все функциональные слова «без собственного лексического значения» (предлоги, подчинительные союзы, знаки препинания и пр.) становятся атрибутами при оставшихся узлах. Полнозначные узлы «аннотируются» ролью в предложении (которая называется «функтором»). Функторов примерно 60: актант, пациенс, адресат, источник, эффект. Учитываются также разные типы пространственных, временных и иных признаков: средство, способ, степень, последствие, условие и др.

Большинство функторов приписывается вручную. Затем создаётся обучающийся модуль, который часть функторов строит автоматически, опираясь на правила и словарные данные, извлечённые из уже аннотированной части корпуса.

К 2002 г. из текстов текущей версии Чешского корпуса в 100 млн. слов проаннотировано в терминах ATS 100 тыс. предложений, средствами TGTS — 20 тыс. предложений; из них 2 тыс. предложений снабжены пометами о коммуникативной структуре (Topic-Focus Articulation — TFA). Последние работы чешских лингвистов обогащают глубинные синтаксические структуры ещё одним видом информации — введением кореферентных связей для личных и указательных местоимений [Kucova, Hajicova, 2004].

Данный уровневый подход к аннотированию текстовых корпусов принят, в основном, в русской и чешской школах КЛ. Он сближается с методами полного лингвистического анализа текстов в системах автоматического понимания текстов (АПТ). Вместе с тем он требует больших трудозатрат со стороны лингвистов, корректирующих результаты автоматических операций.

ТК — это источник различного типа знаний. Информация, содержащаяся в текстовых массивах, без лингвистической обработки не может быть использована. Для извлечения знания требуются мощные лингвистические технологии. Перед корпусной лингвистикой стоят те же проблемы, которые характерны для этапа анализа языковых ресурсов в системах автоматической обработки текстов:

а) сортировка и систематизация текстовых массивов;
б) сегментация текстов;
в) общелингвистический поверхностный анализ, или аннотирование, текстов;
г) внутренняя разметка: расстановка морфологических, синтаксических и семантических обозначений («тэгов»).

Чтобы быть полезным объектом для разных специалистов и чтобы предоставить лингвисту возможность выбрать или собрать нужный ему массив, КЛ систематизирует коллекции текстов — по эпохам, языкам, жанрам, стилям, тематике и т.п. Кодирование метаинформации о тексте документа и его внешних параметрах опирается в большой мере на уже разработанные технологии. Используются разные системы кодирования текстов (HTML, XML и особенно TEI: Text Encoding and Interchange); в частности, систематизация указанных выше русских корпусов основана на стандарте TEI. Этому, а также истории и полезным параметрам КЛ посвящены работы С.А. Шарова и указанная в них литература [Шаров, 2003].

Для системы COBUILD была разработана сегментация: «лёгкая» (выделяются заголовки и подзаголовки текстов) и «нежёсткая» (уточняются или снимаются различного рода пометы к текстам). Сегментация текста — процесс корпусного анализа, при котором части текста делятся сначала на предложения (или словосочетания), а затем вычленяются более мелкие единицы, например, обозначения дат, денежных сумм, названия компаний, адреса, номера телефонов и т.д. При этом синтаксический препроцессор объединяет их в группы непосредственных составляющих по заданным комбинациям признаков.

Грамматики в системе COBUILD построены по принципу *data-driven* («под управлением данных»). Данный принцип противопоставлен принципу *data-based*, когда лингвист задаёт грамматику интуитивно, а корпус используется для проверки её правильности и для подбора примеров. В грамматике *data-driven* существенна лексическая компонента: нет независимого выбора грамматических конструкций и подстановки в них лексем — они работают вместе, создавая определённое значение. Есть списки лексем, для которых характерна определённая «схема», например Vn that; V+C (verb + complement), V + O + A (Verb + Object + Adjunct).

Схемы в такой корпусной лингвистике не правила, а некое обобщение употреблений. В них не различаются синтаксис и лексика (нет такого формального автономного синтаксиса, категориями которого можно было бы манипулировать без обращения к значениям слов [Barlow, 1996]): VP [lose [POSS way]], VP [lose [REFL]], VP [let NP go], VP [let [REFL] go].

Схемы могут быть вложенными. Кроме того, они могут быть связаны с типом дискурса (с учётом включения говорящего и слушающего). В традиционной КЛ нет уровня автономного синтаксиса. Не различаются глубинный и поверхностный уровни синтаксиса. Не проводится различие между категориями «Лексика» и «Структура». Вместо этого имеется формальная часть «Схема — Значение», «Структура — Лексика». Соответственно, и поисковый аппарат в корпусах принимает структуры, состоящие частично из лексем, частично из «тэгов» (грамматических и других помет).

Специалисты, работающие в области корпусной лингвистики, как правило, руководствуются определённой концепцией. Так, например, по мнению В.Тойберта, КЛ противопоставляется когнитивной лингвистике [Teubert, 2001]. Для КЛ не существует «языка мысли». В.Тойберт отрицает всякие репрезентации, ментальные языки, атомы смысла и пр. как нечто нематериальное, символы, абстракции, которые нельзя интерпретировать. По его

мнению, ни в искусственном интеллекте (ИИ), ни в машинном переводе (МП), по сути, нет никакого когнитивного подхода.

КЛ не волнует истинность высказываний. Неважно также, что думает кто-то о реальной воде: слово «вода» означает то, чем и является вода. КЛ имеет дело с языком как социальным явлением. Значение — в словах и текстах. КЛ не интересуют значения изолированных слов вне релевантных для них контекстов. Цитата даёт больше, чем словарная дефиниция слова. Значение неотделимо от формы. Различие в значении всегда сопровождается различием в форме. Корпусный анализ может помочь определить образцы этих форм.

По мнению В.Тойберта, КЛ отказывается от всех теоретических достижений лингвистики после Ф. де Соссюра. В основном, это относится ко всем вариантам порождающей грамматики школы Н. Хомского и его последователей. Исключение составляет аппарат категориальной грамматики [Teubert, 2001]. Универсальная грамматика описывает только ядро языка и ничего не говорит о периферийных зонах, тогда как исследователи языка и изучающие язык нуждаются именно в конкретном материале разных синтаксических конструкций [Barlow, 1996].

Поскольку КЛ интересует не отдельное слово, а текстовые сегменты, разница между лингвистическими и энциклопедическими знаниями размывается. Так, если немецкое слово Machtergreifung означает просто захват власти какой-то группой, ранее исключённой из политической жизни, своими силами, недемократически, то сегмент braune Machtergreifung im Jahre 1933 безоговорочно означает захват власти нацистами. Объясняется это тем, что часто в разных контекстах они заменяли друг друга, были парафразами или анафорически связанными сегментами. Энциклопедическое знание — это не что иное, как дискурсивное знание. Нет значения вне языка, вне дискурса.

КЛ связана с проблемами автоматической обработки текста и, конкретно, с системами автоматического распознавания и понимания текстов [Потапова, 1997; Потапова, 2005] по нескольким признакам:

а) системы АПТ нужны именно для работы с большими массивами текстов. Чтобы добиться каких-то полезных результатов в работающей системе, необходимо знать и учитывать все свойства этих новых для лингвистики объектов — текста как целого и массива текстов. КЛ формирует, исследует и описывает их как информационный ресурс;

б) технологии и приёмы первичной обработки «сырых», непрепарированных текстов в прикладных системах (машинный перевод и другие системы АПТ) во многом совпадают с теми, которые приняты или отрабатываются в КЛ. Так, чтобы создать параллельный корпус, нужны алгоритмы и программы сегментирования текста на такие (значимые) единицы, которые могут быть сопоставлены друг другу;

в) массивы КЛ — это надёжный источник формирования словарей, в том числе двуязычных, и выуживания информации, которую надо включать в словарную статью (иллюстрации значений слова, сведения об актантной структуре слова и др.); это источник создания конкордансов, словников, тезаурусов и других инструментов, необходимых для автоматического анализа произвольных текстов. Составление словарей — одно-, дву- или многоязычных — должно подтверждаться массивами КЛ, если не полностью базироваться на них [Леонтьева, 2006].

Тем самым КЛ не исключает, а дополняет традиционную лингвистику, становится опорой общей лексикографии. Ведь лексикография работает не только с простыми единица-

ми и их контекстом, но и с большими текстовыми сегментами, единицы которых определены на лексическом и синтаксическом, включая порядок слов, уровнях (многословные единицы, термины, коллокации, обороты). Традиционная лингвистика всё больше нуждается в более крупных, чем слово, единицах и в обосновании их выделения обращением к КЛ; она тяготеет к изучению семантической связности (lexical solidarities, collocations, set phrases, valencies, case roles, thematic roles, semantic frames and scripts). КЛ проясняет понятие текстового сегмента эвристическим определением семантической связности: совместной встречаемостью схем (цепочек), которые тем самым связаны определёнными семантическими отношениями.

Статистика совместной встречаемости и явное выражение шаблонов (комбинации квантитативных и категориальных признаков) позволяют изучать «размытые» значения (fuzzy meanings). КЛ допускает втягивать пользователя в дискурс и включать его определения в универсум цитат и контекстов [International Journal of Corpus Linguistics 1996 и др.].

Эмпирической базой многоязычной корпусной лингвистики служит (виртуальный) массив всех текстов, когда-либо переведённых на другой язык, вместе со своими переводами. Теоретическая основа та же, что и для одноязычных корпусов, т.е. значением текстовой единицы считается парафраза, а полное значение текстового сегмента в этом дискурсивном универсуме заключено в истории (сумме) всех переводных эквивалентов данного сегмента.

Создание параллельных и многоязычных корпусов столкнулось с трудной задачей «выравнивания», т.е. разбиения параллельных текстов на единицы, которые можно сопоставить друг с другом.

Большинство программ выравнивания в параллельных корпусах основывается на том, что в переводе сохраняются те же границы предложений и абзацев, что и в исходном тексте. В действительности же разные типы текстов требуют перестановки или сокращения (например, в юридических текстах) числа предложений. Пословные соответствия (предлог — отсутствие предлога, падеж — предложная конструкция) составляют незначительное число. Минимальные единицы перевода могут состоять из одного слова или нескольких слов, переводимых как целое, а не пословно. Переводные эквиваленты соответствуют текстовым сегментам одноязычного корпуса. Значение единицы перевода содержится в её переводных эквивалентах на другие языки. Идентификация единиц перевода требует интерпретации: единый это эквивалент или комбинация нескольких. Текстовый сегмент является единицей перевода по отношению только к тем языкам, в которых он переводится как единое целое. Неоднозначные единицы перевода имеют столько значений, сколько есть несинонимичных переводных эквивалентов. Данная единица перевода языка А может иметь два несинонимичных эквивалента в языке В и три — в языке С. Объявить какие-то эквиваленты синонимами — это акт интерпретации; сначала надо понять текст, а это компьютерам недоступно. Практическое использование корпусной лингвистики — помощь переводчику путём обработки параллельных массивов. Последние — это хранилища переводов. Использовать их гораздо более эффективно, чем традиционные двуязычные словари, особенно

если в массиве учтены жанр и тип текстов: выбирается тот эквивалент, контекстная проекция которого больше всего совпадает с профилем текстового сегмента.

Анализ «по образцу», или прецедентный анализ, важен не только для систем МП, как отмечалось ещё в ранних работах по МП, но и как серьёзное подспорье при анализе свободных текстов. И всё же проблема формирования параллельных корпусов достаточно трудна — и не только содержательно, но и чисто технически. С одной стороны, нужно сделать эксплицитной всю релевантную информацию. С другой стороны, текст, отягощённый тэгами, становится нечитабельным. Любые изменения в размеченном корпусе — всегда проблема.

Многие сторонники КЛ считают, что для обработки многоязычных массивов текстов продуктивно использовать языково-независимые подходы [Greenstette, Segond, 1997]. В RXRC (Ranc Xerox Research Centre) создано несколько средств АОТ, работающих на основе автоматов с конечным числом состояний и трансдукторов (The transducer is a finite-state machine which consumes input while producing output). Эти простые методы обработки оказались применимы к очень большому количеству лингвистических структур.

Разработанные средства были использованы в нескольких прикладных задачах: задаче извлечения терминологии (information extraction), в системе помощи переводчику и в информационном поиске (cross-language information retrieval). Технология автоматов с конечным числом состояний имеет много достоинств: это хорошо изученные механизмы, поддающиеся разным математическим операциям, их можно по-разному комбинировать, вставлять в другие процедуры и т.д. Правила трансформаций могут включать контекст, тем самым не требуя специальных программных решений. Модульность и возможность включать контекстные условия в структуру данных позволяют быстро приспосабливать подобные пакеты (suits) АОТ к другим языкам. Пакеты включают языково-независимые правила сегментации (tokenizer), морфологические анализаторы, программы построения гипотез для неузнанных слов, программы приписывания частей речи (POS: part-of-speech taggers) и программы сборки именных групп (noun-phrases extractors). Такие пакеты созданы в RXRC для семи европейских языков, готовятся ещё для семи (русского, чешского, венгерского и др.).

Главное в подходе RXRC — разработка надёжных и всё более мощных технических решений, применимых к любым массивам текстов на естественном языке.

В настоящее время результаты корпусных исследований находят основное практическое применение в создании больших контекстно-ориентированных тезаурусов, которые увеличивают семантическую силу при работе систем информационного поиска. Так, в системе ACRONYM (Automated Collocational Retrieval of «Nyms») собираются концептуально родственные единицы, называемые Nyms («нимы», по аналогии с синонимами и др.) [Collier, Pacey, Renouf, 1998]. При этом не проводится никакая предварительная лингвистическая разметка (считается, что это слишком «дорогой» процесс на очень больших массивах), кроме перевода числовых цепочек в обобщённые категории. Проводится кластерный анализ, вычисляется мера подобия соответственно правых и левых контекстов для выделенных единиц (слов и словосочетаний), учитывается частота появления сходных контекстов и т. п. Сначала собираются группы родственных слов (нимов) первого порядка, что уже может хорошо работать для информационного поиска, затем рядом уточняющих процедур строятся нимы второго порядка, которые должны удовлетворить и лингвистов. Приведём пример построенного в системе ACRONYM списка нетривиальных «родственников» для четырёх английских слов:

Таблица 1. Пример организации данных в системе ACRONYM

Node	Nyms
Key	crucial important vital significant essential main fundamental major strategic specific
Medicine	medical medicines sciences mathematics biology science chemistry psychology physics clinical
Pretty	fairly quite incredibly extremely terribly really nice extraordinarily lovely sexy
Testing	tests test tested assessment monitoring screening research rigorous clinical curriculum

Таким образом, текстовый корпус — это особый, совершенно новый тип словесного единства. Можно выделить четыре базовых качества, делающих собрание текстов корпусом [Рыков, 2003]:
– расположение на магнитном носителе;
– процедуры отбора материала, обеспечивающие его репрезентативность;
– единство разметки на носителе;
– конечный размер.

Возможно формирование не только универсальных, т.е. представительных с учётом разных жанров для всего языка, но и специализированных (для каких-то задач) корпусов текстов (к их числу относятся КТ звучащей речи).

Создание устно-речевых баз данных (УРБД) является на сегодняшний день первоочередной задачей в свете актуальности проблемы автоматизации процессов распознавания и понимания речи, идентификации говорящего по голосу и речи, синтеза речи, устного перевода. В современном мире быстро развивающихся информационных технологий, когда «умные» дома с управляемыми голосом приборами стали реальностью, необходим «строительный материал» для подобных систем. Одними из таких «кирпичиков» и являются фонетические базы данных, или УРБД. Формирование репрезентативных УРБД является одним из условий успешного решения прикладных задач.

Формирование УРБД многоцелевого назначения применительно к различным языкам мира является одной из приоритетных задач современного речеведения [Потапова, 1997].

Подавляющее большинство конструируемых сегодня автоматизированных систем, работающих со звучащей речью, так или иначе используют устно-речевые базы данных. В частности, УРБД находят применение там, где используются вероятностные и статистические методы анализа и синтеза речевого сигнала. В первую очередь здесь следует упомянуть системы автоматического распознавания и синтеза речи, идентификации и верификации говорящего по голосу и речи, идентификации психофизического и эмоционального состояния говорящего по речи, а также обучающие системы. Далее, УРБД составляют основу автоматизированных систем, в задачи которых входит сбор и хранение речевых сообщений, поиск и выдача записанных речевых сообщений по запросу (например, автоматизированные системы приёма голосовых сообщений в колл-центрах, комплексы для тестирования каналов связи). В ряде других случаев использование УРБД, не будучи строго необходимым технически, оказывается разумной альтернативой разработке сложных процедурных решений.

Как правило, УРБД содержат большие объёмы численной информации, трудно поддающейся автоматическому структурированию и сжатию. В то же время в силу специфики систем, в которых применяются УРБД, в большинстве случаев эта информация должна быть доступна для обработки в режимах, близких к режиму реального времени. Поэтому структура УРБД должна обеспечивать максимальное быстродействие системы при разумной ресурсоёмкости. По причине большого объёма информации изменение, а следовательно, и оптимизация структуры действующей УРБД обычно является технически трудновыполнимой и крайне нежелательной операцией. С учётом многообразия задач, для решения которых применяются УРБД, это означает, что её структура должна быть универсальной и, как следствие, максимально простой.

При разработке УРБД неминуемо встаёт проблема выбора системы управления базами данных (СУБД) [Белолипецкий, Буря, 2004]. Здесь возможны следующие варианты: выбрать существующую хорошо зарекомендовавшую себя СУБД из числа присутствующих на рынке информационных технологий или разработать свою СУБД специально для этой задачи.

При выборе из возможных вариантов разработчики руководствуются обычно следующими требованиями к СУБД, на которой реализуется речевая база данных:

— СУБД должна осуществлять удобное хранение как обычного текста, так и больших бинарных данных (BLOB). Под удобством понимается простота программного интерфейса для извлечения и записи данных;

— СУБД должна поддерживать хранение данных большого объёма (и большого суммарного объёма). Объём речевой БД может в несколько раз превышать объём доступной оперативной памяти;

— СУБД должно обеспечивать достаточно быстрый доступ к данным в режиме чтения. Это требование вступает в противоречие с предыдущим требованием, поэтому потребуется выработать некоторый компромисс или определить, какое из этих требований является приоритетным. Следует также учитывать специфику работы с речевой БД. Наиболее распространённым является сценарий последовательного перебора всех речевых образов (т.е. их последовательного чтения из БД), поэтому обширное кэширование данных не приведёт к значительному росту производительности;

— СУБД должна иметь средства для различных статистических расчётов по текущему состоянию речевой БД;

— СУБД должна иметь средства для импорта и экспорта речевых бинарных данных (а также сопутствующих им текстовой информации);

— СУБД должна предоставлять возможность как однократного полнообъёмного прочтения бинарного речевого образа, так и поэтапного (постраничного) чтения. Это требование возникает из-за различных потребностей алгоритмов обработки данных, алгоритмов распознавания и процедур импорта/экспорта. В реляционных СУБД существует подобный механизм, позволяющий указывать оптимизатору запросов, требуется ли последовательное получение результатов запроса и однократное (спецификаторы FIRST_ROWS, ALL_ROWS);

— СУБД должна позволять получить речевые данные на внешнем носителе информации в стандартном формате звукозаписи для возможности использования большинства программ обработки звука. Это требование можно обеспечить либо постоянным хранением бинарных данных в файлах со стандартным форматом (наподобие типа данных BFILE в Oracle), либо с помощью развитых средств импорта/экспорта.

Желательно также предусмотреть средства, облегчающие (автоматизирующие) пакетный запуск алгоритмов различных видов обработки речевых данных.

Существующие реляционные СУБД мало подходят для хранения речевых данных; наиболее удобный механизм реляционных СУБД, позволяющий хранить речевые данные, – тип данных BLOB. К сожалению, использование этого типа данных для хранения речевых данных сопряжено с рядом проблем. Тип данных BLOB рассматривается разработчиками реляционных СУБД как дополнительный или даже необязательный. Вследствие этого операции для работы с ним малочисленны и неоптимизированны. Во многих реляционных СУБД нет таких операций для работы с типом данных BLOB, как вставка и удаление части данных, есть только полная перезапись и обнуление. Уже одно то, что в языке SQL нет никаких операций для работы с типом данных BLOB, показывает, насколько затруднено его использование.

У ряда реляционных СУБД есть возможность хранить данные во внешних файлах (например, тип данных BFILE в СУБД Oracle), что позволяет использовать средства файловой системы для работы с данными. Но и это не панацея. С тем же успехом можно и не пользоваться СУБД, а хранить речевые данные просто в файлах.

Несколько лучше отвечают требованиям речевой базы данных объектно-ориентированные СУБД (ООСУБД). Особенно заманчиво использовать для подобной цели ООСУБД типа Jasmine, ориентированные на разработку мультимедийных приложений. В этом случае не возникает проблемы с надлежащим типом данных, поскольку ООСУБД так или иначе предоставляют возможность создать свой тип данных (класс), для которого можно задать формат хранимых данных, описать нужные операции по обработке данных. Также мультимедийные ООСУБД работают с большими объёмами данных намного эффективнее реляционных.

Интерес к созданию корпусов звучащей речи был в значительной степени инициирован разработками в области автоматического распознавания речи, где исследователям приходится сталкиваться с акустической вариативностью звуковых единиц языка, которая имеет разнообразные причины: от системной контекстной вариативности, обусловленной коартикуляцией, до психофизиологического и эмоционального состояния говорящего, а также технических характеристик микрофона, который используется при записи речевого материала [Кривнова, Захаров, Строкин, 2001; Кривнова, 2008][1].

Первые речевые корпусы появились в середине 80-х годов в США, где их разработка финансировалась прежде всего Министерством обороны. При поддержке этого ведомства были созданы: TI-DIGITS корпус (1984 г.) для тестирования систем распознавания изолированных цифр и цифровых последовательностей; Road Rally для анализа и распознавания ключевых слов (word spotting) и King Corpus для систем идентификации говорящего (speaker recognition). В рамках государственной программы развития лингвистических технологий, известной как ARPA/DARPA (Advanced Research Projects Agency), это же министерство финансировало создание уже упоминавшегося выше корпуса TIMIT, который послужил прототипом для многих других речевых баз данных. При этой же финансовой поддержке были разработаны специализированные речевые корпусы Resourse Management

[1] Подробнее об УРБД для русской речи см. работы О. Ф. Кривновой и её соавторов.

(RM) и Wall Street Journal (WSJ) для исследований в области распознавания слитной речи, а также Air Travel Information Service (ATIS) для исследования спонтанной речи и понимания естественного языка в диалоговых системах [Кривнова, Захаров, Строкин, 2001; Кривнова, 2008].

Накопленный к концу 80-х годов опыт показал, что создание представительных речевых корпусов требует кооперативных усилий исследовательских институтов, индустриальных компаний и государственных спонсоров. Финансовые и временные затраты на разработку высококачественных ресурсов оказались очень велики. Эксперты отметили, что дорогостоящие, но необходимые для развития информационных технологий ресурсы не должны разрабатываться для какой-то одной специальной системы или задачи [Godfrey & Zampolli, 1997]. Они пришли к выводу, что ресурсы должны обеспечивать возможность их многократного использования разными пользователями, т. е. быть общедоступными, и более чем для одной цели, т. е. быть многофункциональными. В связи с этими требованиями возникла проблема стандартизации лингвистических описаний, согласования форматов представления информации в разных видах лингвистических ресурсов и их типологии (подробнее см. [Gibbon et al., 1997]).

УРБД разрабатываются для решения конкретной задачи. Круг возможных применений велик, однако конкретная задача задаёт непосредственные характеристики базы.

В 1991 году в США был создан лингвистический консорциум (LDC – Linguistic Data Consortium), который поддерживает создание новых языковых корпусов и распространяет ресурсы, полученные из разных источников. В частности, в настоящее время LDC предлагает речевые корпусы, которые в совокупности содержат многие сотни часов звучащей речи. Технологический центр в штате Орегон (CSLU – Center for Spoken Language Understanding) коллекционирует, аннотирует и распространяет телефонные речевые корпусы. Активность Центра поддерживается индустриальными спонсорами. Собранные корпусы доступны университетам по всему миру бесплатно. Этот центр располагает также многоязычным корпусом для оценки алгоритмов идентификации языка, который состоит из фрагментов спонтанной речи на одиннадцати разных языках мира. В 1995 году координационный центр лингвистических ресурсов (ELRA – European Language Resources Assosiation) был образован и в Европе (более подробные сведения об истории создания и задачах этой ассоциации можно найти, например, в обзорных статьях [Mariani, 1996; Teubert, 1996]). В распоряжении этого центра находятся речевые корпусы для большинства официальных языков Европейского союза: для британского и шотландского вариантов английского языка, голландского, датского, шведского, немецкого, французского, итальянского, испанского, – а также несколько многоязычных корпусов. В настоящее время в результате осуществления программы Copernicus ELRA распространяет также речевые корпусы для языков Восточной Европы (польский, болгарский, эстонский, румынский и венгерский). На сайте Европейской ассоциации в Интернете можно найти предложения и речевых корпусов для русского языка.

Сравнение различных акустических баз данных позволяет сформулировать некоторые обязательные требования к современной фонетической базе данных, предназначенной для фундаментальных и прикладных исследований. УРБД для прикладных исследований, в частности, в области синтеза и распознавания речи, должны обеспечивать решение следующих задач [Скрелин, Щербаков, 2003]:

- Внесение в УРБД звуковых эталонов — оцифрованных записей речи нормативных дикторов в разных стилях речи, от спонтанной речи и чтения текстов, полученных на основе её расшифровок, до чтения списка слов. Другими словами, в БД необходимо включить звуковой материал, представляющий максимальную вариативность реализа-

ции языковых единиц (фонем и интонационных конструкций) в различных условиях речевой деятельности человека.

- Внесение сегментной информации и подробного фонетического описания включаемых звуковых образцов, поскольку необходимо снабдить этот материал подробным описанием: адресами границ звуков и интонационных единиц, словоформ и слогов (так как существуют различные методики распознавания и синтеза речи с учётом базовых единиц), а также фонемной и подробной фонетической транскрипцией.
- Обеспечение эффективного выполнения запросов к содержимому УРБД для поиска нужных звуковых фрагментов по их транскрипционным описаниям и указанным в описаниях признакам.

Недостаточная проработка реализации любого из вышеперечисленных пунктов существенно снижает ценность УРБД в целом [Скрелин, Щербаков, 2003]. В реальности же для создания УРБД можно использовать любой ПК, оснащённый звуковой платой, совместимой с SB16. Производят запись речевого материала, для чего могут либо приглашать дикторов и производить запись в лабораторных условиях, либо собирать материал из широко доступных источников, например, теле- и радиотрансляций, вещаний в интернете и т.п. Для записи Интернет-трансляций нужна программа, которая может выполнять функции магнитофона, и программа, поддерживающая и воспроизводящая формат потокового аудио из интернета. Например, CoolEdit 1.0, которая, выполняя функции записи, является одновременно и звуковым редактором, и заменяет собой Windows Media Player, RealPlayer и т.д. Чтобы облегчить сбор и хранение данных УРБД, разрабатывается специальная оболочка. Она представляет собой отдельную программу-приложение, которая обладает возможностями: записи/воспроизведения фрагментов; хранения информации о фрагменте; хранения информации о дикторе (если такая информация нужна); поиска информации по различным параметрам.

После записи речевого материала, ввода речевого материала в компьютер (оцифровки) и сохранения его, эксперт-фонетист производит транскрибирование материала; файл транскрипции имеет, как правило, формат txt. Затем эксперт-акустик производит сегментацию материала, сохранённого в файлах форматов wav и txt, с его последующим сохранением в две разные папки, поименованные соответственно WAVE и TEXT. Эксперт-фонетист создаёт правила перехода «звук–буква», причём звуки представлены специальным алфавитом, варианты которого создаются для каждого языка. В настоящее время одним из таких стандартов можно назвать фонетический алфавит Sampa (Speech Assessment Methods Phonetic Alphabet). Он представляет собой Международный фонетический алфавит, записанный символами ASCII с соответствующими изменениями под конкретный язык.

Некоторые речевые базы данных для русского языка создавалась в рамках европейских проектов SpeechDat(II) и SpeechDat(E) [http://www.auditech.ru]. Целью проектов, объединенных названием SpeechDat, является создание речевых баз данных в странах Европы посредством записи речи в реальных условиях через телефонный канал стандарта ISDN. Базы данных призваны служить общим ресурсом для 20 европейских языков и диалектов и способствовать разработке общих систем телесервиса.

В проектах SpeechDat, профинансированных Европейским союзом, были представлены крупнейшие промышленные и академические организации. Все базы данных, созданные в рамках этих проектов, имеют стандартный дизайн и прошли все этапы валидации.

Созданные в рамках проектов SpeechDat речевые базы данных удовлетворяют следующим требованиям:

- охватывают фонетически репрезентативные слова, слова-команды, словосочетания, числа, цифры, числовые последовательности, фонетически репрезентативные предложения;
- представляют различные стили произнесения (команды, речь-чтение и спонтанная речь);
- фиксируют окружающую акустическую обстановку;
- пригодны для разработки и обучения надёжных систем распознавания речи для телесервисов.

В речевой базе данных SpeechDat(II) представлено 48, а в базе данных SpeechDat(E) — 50 слов и выражений, как *СПОНТАННО ПРОИЗНЕСЁННЫХ*, так и *ПРОЧИТАННЫХ*. Продолжительность записи (диалога между диктором и компьютером) составляла 8–10 минут в зависимости от темпа речи.

Исходный словарь базы данных содержит списки наиболее употребительных слов и команд из компьютерной лексики, цифр и цифровых последовательностей, названий крупных городов и фирм, обозначающих время фраз, дат, денежных единиц, телефонных номеров, номеров кредитных карт, сочетаний «имя-фамилия», фонетически богатых слов и предложений, а также спеллинг (побуквенное произнесение) слов.

Технические характеристики записывающей установки были стандартизированы для всех речевых баз данных. Записи проводились в автоматическом режиме через реальный цифровой телефонный канал европейского стандарта ISDN. Сигнал имел формат: 8 бит, 8 кГц, A-закон. Качество соединения и линии связи характеризовалось отношением сигнал/шум. Непригодные по зашумлённости записи исключались.

Обработка речевого материала выполнялась экспертами по речевой акустике. Она заключалась в многократном прослушивании всех звуковых файлов и их аннотации, которая производилась в соответствии со спецификацией, разработанной для участников проекта SpeechDat(II).

Аннотация подразумевала внесение следующей информации в файл-метку:

- орфографическая запись высказывания;
- специальные пометки, указывающие на наличие возможных шумов, оговорок, обрывов записи;
- оценка качества записи;
- данные о дикторе (возраст, пол, региональный акцент);
- тип телефонного аппарата;
- тип акустического окружения.

Из всех слов, произнесённых дикторами разборчиво и без оговорок, был составлен лексикон (файл LEXICON) с указанием частоты встречаемости каждого слова и его фонематической транскрипции. Часть слов приведена с вариантами произнесения (разговорный вариант).

Полученный лексикон насчитывает около 16500 единиц. Фонематическая транскрипция лексикона выполнена в соответствии с системой символов Russian SAMPA (машинно-ориентированного языка). Кроме этого имеется файл акустического качества каждого речевого сигнала, файл информации о респонденте (пол, возраст, регионально-диалектальная принадлежность), файл содержимого базы данных.

Файл DISIGN содержит полное описание базы, её словаря, записывающей платформы, полную информацию о лексиконе (особенности произношения, частота встречаемости фонем и др.).

Поддержание стандартов качества созданных баз данных обеспечено двумя ступенями валидации, которая выполняется фирмой SPEX (Speech Processing Expertise Centre), созданной в рамках проекта SpeechDat для проверки качества и соответствия стандартам созданных баз данных.

В течение многих лет ведутся активные разработки в области формирования многоязычных УРБД на кафедре прикладной и экспериментальной лингвистки Московского государственного лингвистического университета (МГЛУ). При этом охвачены различные языки, включая языки этнических меньшинств Российской Федерации. В разработках участвует ряд языковых кафедр МГЛУ. В качестве примера приведём некоторые из УРБД.

УРБД для французского языка разрабатывалась на кафедре прикладной и экспериментальной лингвистики МГЛУ и в Центре фундаментального и прикладного речеведения МГЛУ (директор Центра — Р.К. Потапова) в рамках проекта «Корпусная лингвистика многоцелевого назначения».

В задачу входило формирование фонетической базы данных французского языка, представленной звучащими текстами. Первой задачей создания УРБД была разработка свода правил соотношения «звук-буква» для французского языка.

Правила были сведены в таблицы, которые включали рубрики: звук, буква/буквосочетание, примеры, примечание. В примечании давались исключения из правил и дополнительная информация.

При транскрибировании использовались фонетические шрифты: NewtonPhoneticNt, Phonetic TM, Phonetic TMUniv, WP Phonetic. Помимо этих правил в тот же корпус вошли таблицы французских гласных, согласных, полугласных, носовых гласных, а также таблица используемых в базе транскрипционных значков международного фонетического алфавита.

Первичный корпус базы представлен фрагментами французской речи, подлежащей сегментации дофразового и фонемного уровней (в зависимости от внутренней спецификации задачи). Записи проводились с помощью программ Cool Edit 2000 и Real Player Plus 8.0. Ряд записей представляет собой оцифрованные фонограммы текстов разного характера, полученные с использованием материала на аудиокассетах (условия оцифровки: 22050 Гц, 16 бит, моно). Тексты включали монологи, диалоги, полилоги, отрывки из театральных спектаклей и др. в исполнении 25 мужчин и 20 женщин. Общее время — 15 час.

Другие записи представляли собой сообщения новостей, взятые с разных порталов Интернета в прочтении 25 мужчин и 20 женщин. При записи новостей он-лайн возникли некоторые трудности: при загрузке файлом реального времени .rm происходили изменения бит-рейта, которые отразились на качестве звучания речи. Последующая стадия обработки звука позволяла компенсировать этот недостаток. После записи речь в файлах подлежала сегментации и помещению в отдельные файлы, соответствующие определённым сегментам.

Далее проводилось аннотирование. Для ряда текстов имелись и видеозаписи, что существенно расширило базу данных и послужило основой для последующих разработок в области создания мультимодальных БД.

В задачу **УРБД для арабского языка** входило формирование фонетической базы данных арабского языка, представленной звучащими текстами. База данных разрабатывалась в том же Центре в рамках проекта «Корпусная лингвистика многоцелевого назначения». Первой задачей создания УРБД была разработка корпуса правил соотношения «звук-буква» для арабского языка.

Помимо правил, база содержала: папку с файлами текста (txt); папку со звуковыми файлами (wav); папку со звуковыми файлами несегментированного материала (тренировочный комплекс).

Исходным материалом служили файлы, представленные в качестве примера в следующей таблице.

Таблица 2

Исходные файлы

№	Файл	Время звучания**	Дикторы	Источники звучащей речи
1	1_a	45:17	7m,1f	Aljazeera
2	1_b	22:46	5m	Aljazeera
3	2_a	30:10	8m,3f	Aljazeera
4	2_b	14:16	7m,	Aljazeera
5*	3_a	45:34	4m,3f	London course of Arabic
6*	3_b	15:31	4m,3f	London course of Arabic
7*	4_a	45:05	4m,3f	London course of Arabic
8*	4_b	11:28	4m,3f	London course of Arabic
9*	5_a	45:32	4m,3f	London course of Arabic
10*	5_b	13:04	4m,3f	London course of Arabic
11*	6_a	46:09	4m,3f	London course of Arabic
12*	6_b	12:37	4m,3f	London course of Arabic
13***	7_a	45:21	3f	Alarabia
14***	7_b	45:09	3f	Alarabia
15****	8_a	46:08	10 m,6f	Aljazeera
16****	8_b	28:16	10m,6f	Aljazeera

Общее время звучания — 8,5 часов. Дикторы — 59 мужчин и 20 женщин.

Ниже представлены примеры оцифрованных записей (22050 Гц, 16 бит, моно) с аудиокассет. Все файлы типа Windows PCM (wav).

Таблица 3

№	Файл	Время звучания	Дикторы	Источники звучащей речи
1	Aljazeera1	13:42	5m	Aljazeera
2	Bbc1	9:48	2m, 1f	BBC
3	Dw2	30:19	7m,1f	Deutsche Welle
4	Jaber Ibn Hayan	27:26	4m,2f	VOA

Общий объём звучащего материала УРБД составил 1 Гб (соответствует 6,5 ч звучания). Представлены голоса 79 дикторов (59 мужчин и 20 женщин) — носителей различных произносительных вариантов арабского языка. УРБД включает 2070 пар файлов (аудиоматериал/текст в орфографии и транскрипции). Аудиофайл представляет собой оцифрованную запись фрагмента арабской речи в формате WAV (Microsoft Wave). Такое представление данных позволяет легко проводить поиск и сопоставление, а также инкорпорировать информацию в любые автоматизированные речевые системы, что соответствует задаче формирования УРБД многоцелевого назначения.

В ряде случаев в качестве исходного материала использовались файлы сжатых аудиоформатов (в частности, WMA-8, качество FM Radio), которые в дальнейшем были также приведены к формату WAV с указанными выше параметрами. Максимальная длительность звучания одного фрагмента — 81,8 с., минимальная — 0,2 с. Каждый фрагмент включает голос одного диктора, записанный в одних и тех же условиях.

На рис. 1 представлено распределение значений длительности сегментов в УРБД (группированный ряд).

Формирование УРБД арабского языка осуществлялось по следующим этапам:

1. Проводилось многократное прослушивание звучащего материала в полном объёме аудиторами — специалистами в области арабского языка, а также аудиторами – специалистами в области экспериментальной фонетики.
2. Транскрибировался каждый прослушанный звучащий файл с использованием системы транскрипции, принятой в международной информационной сети (SAMPA for Arabic).
3. В процессе аудитивного анализа и транскрибирования использовались скомпонованные на 1-м этапе данного исследования файлы (N = 16), содержащие аутентичный арабский материал в исполнении дикторов — мужчин и женщин.
4. В ходе аудитивного анализа отбраковывалась часть материала вследствие наличия зашумлённости речевого сигнала. Итоговое время звучания речевого материала составило 8 часов.
5. Параллельно проводилась сегментация двух видов: акустическая (с использованием программ CoolEdit 1.0 и Sound Forge 4.5 c) и текстовая (с использованием текстового процессора Microsoft Word).

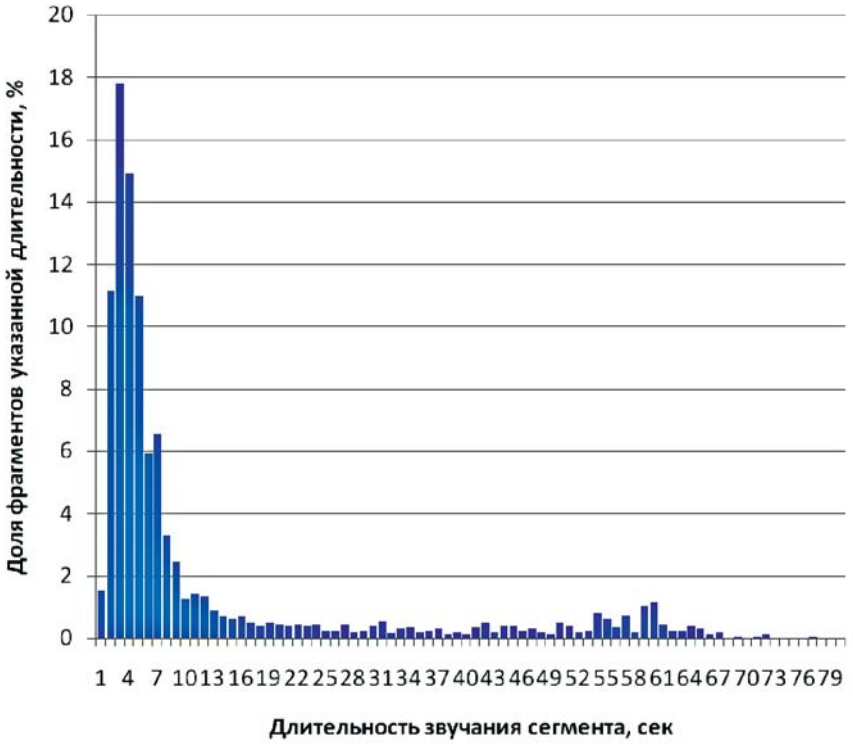


Рис. 1. Распределение значений длительности сегментов УРБД арабского языка

6. Полученные речевые сегменты (nΣ = 2070) подвергались повторному аудитивному анализу с целью подтверждения точности соответствия акустической и текстовой информации в рамках каждого сегмента.
7. Отсегментированный материал (звуковые и текстовые файлы) записывались на оптические носители информации (компакт-диски).

Среди устно-речевых фрагментов, представленных в УРБД, преобладают реплики диалогов (спонтанных или разученных), которые в совокупности составляют около 75% выборки. В основном этим объясняется доминирование в УРБД сегментов длительностью до 10 с. Примеры фрагментов представлены на рис. 2, 3. Оставшаяся часть материала представляет собой фрагменты монологической спонтанной речи (10%) или записи профессионального чтения текстов на литературном арабском языке (15%).

При сегментации больших участков, содержащих голос одного диктора, на фрагменты длительностью до полутора минут (в соответствии с ограничениями, поставленными при разработке структурной концепции УРБД) во всех случаях соблюдалось правило, согласно которому граница фрагментов не должна разрывать предложение (или дыхательную группу) в речевом потоке. Таким образом, фрагменты, длительность которых превышает 10 с., могут содержать от 1 до 4 единиц этого типа (что обусловливает наличие неярко выраженных максимумов на гистограмме в районе 30, 45 и 60 с.).

Все фрагменты, включенные в УРБД, затранскрибированы с использованием международного универсального фонетического алфавита SAMPA (см. http://www.phon.ucl.ac.uk/home/sampa/home.htm, а также http://en.wikipedia.org/wiki/SAMPA) с указанием долго-

ты гласных. В транскрипционную запись была также введена некоторая информация о грамматико-морфологической структуре слов (показано наличие артиклей, подвергшихся фонетической ассимиляции), поскольку эта информация может быть полезной в дальнейшем при использовании УРБД в целях изучения фонетической вариативности арабской речи. Транскрипция помещена в текстовые файлы (*.TXT), имена которых идентичны именам соответствующих аудиофайлов (*.WAV). Данная УРБД охватывает различные региональные и гендерные произносительные варианты арабского языка, а также различные виды речевой деятельности.

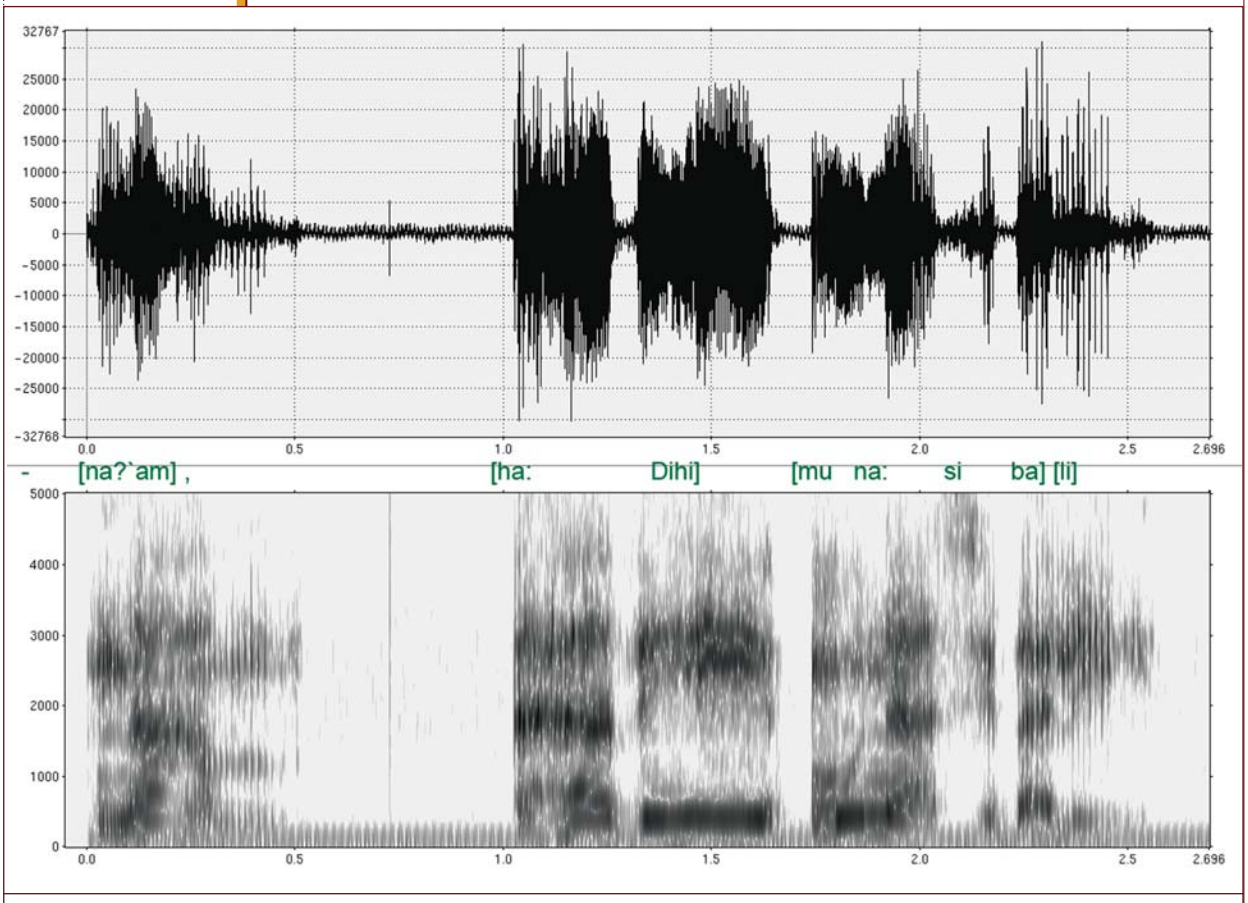


Рис. 2. Пример фрагмента УРБД (5B4MB-12.WAV, TXT), мужской голос

Аналогичные УРБД разрабатывались в Центре фундаментального и прикладного речеведения для турецкого, китайского языков и некоторых языков этнических меньшинств РФ.

Центры цифровых баз данных (les centres de ressources numйriques, CRN) созданы по совместной инициативе Управления научной информации и научного отдела «Человек и общество» Национального центра научных исследований Франции (Centre Nacional de la Recherche Scientifique, CNRS). Центр баз данных для устной речи (Centre de Ressources sur la Description de l'Oral, CRDO) и центры цифровых баз данных (CRN) сосредоточили

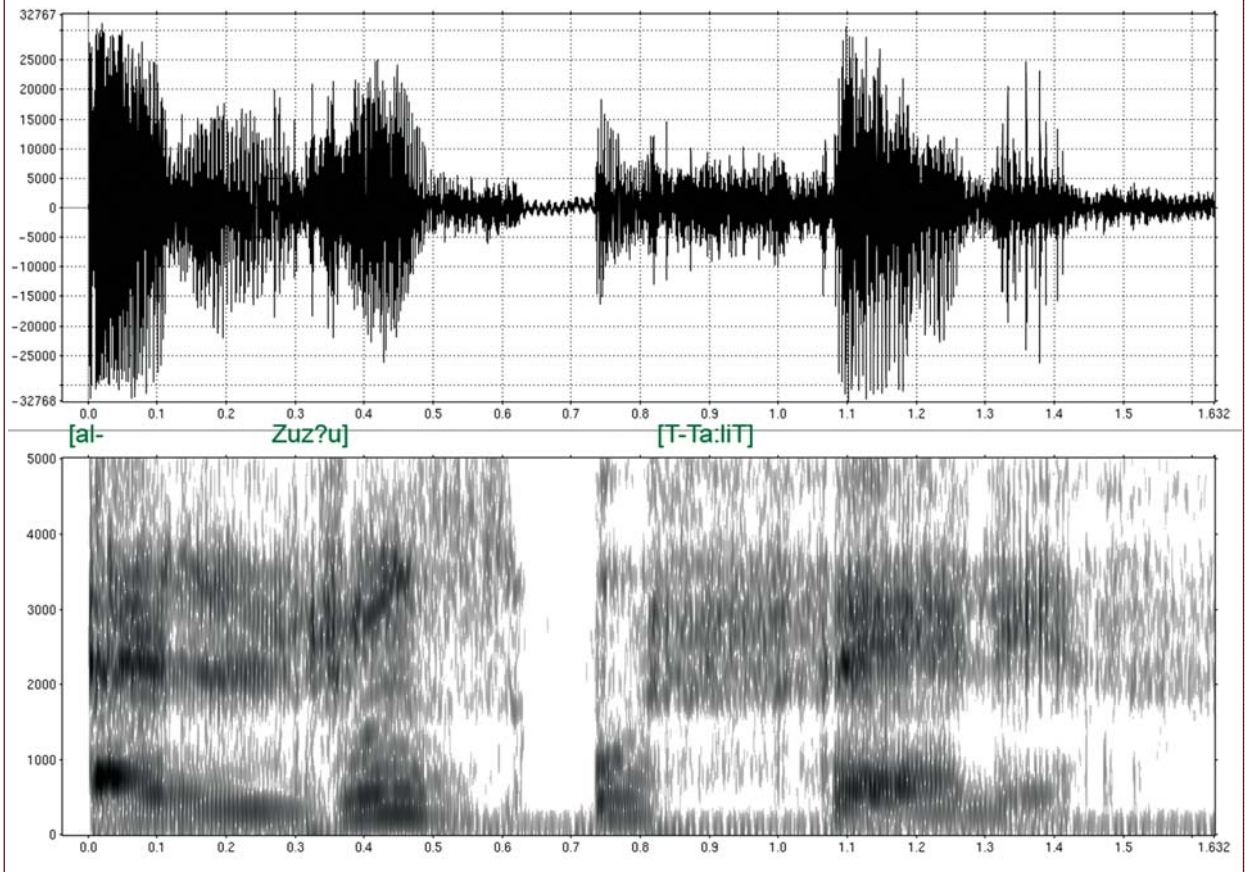


Рис. 3. Пример фрагмента УРБД (5B6F-1.WAV, TXT), женский голос

своё внимание на ресурсах устной речи. В 2006 г. Национальный центр научных исследований поручил Лаборатории языка и речи и Лаборатории языков и цивилизаций с опорой на устную традицию сформировать Центр баз данных для устной речи (CRDO) для таких задач, как каталогизация имеющихся массивов, централизация и обеспечение доступа к ресурсам и инструментам для изучения устной речи. Одним из главнейших компонентов этой работы стало формирование УРБД для различных языков мира.

В 2006–2007 гг. работы, проводимые CRDO, включали решение следующих задач:

— создание и отладка серверной структуры для хранения и обеспечения доступа к ресурсам УРБД;

— разработка структуры метаданных, описывающих содержание каждой УРБД с использованием лингвистических дескрипторов, соответствующих международным стандартам;

— обеспечение авторизованного доступа к УРБД для участников проекта CRDO;

— обеспечение возможности совместного доступа к УРБД, редактирования метаданных и другой информации [Bel, Blache, 2006: 13–14].

Создание серверной структуры проводилось в два этапа. На первом этапе была разработана структура реляционной базы данных для хранения и непосредственного редактирования метаданных. Эта реляционная УРБД была также призвана стать связующим звеном для всех будущих УРБД, формируемых CRDO.

При формировании структуры УРБД учитывались следующие требования: 1) структура УРБД не должна исключать любую априорную информацию; 2) данные, выдаваемые УРБД по запросу, должны быть структурированы согласно международным стандартам.

Информация об организациях, создающих УРБД, хранится в независимой базе данных (617 учреждений, см. http://teck.lpl.univ-aix.fr/institution/institution-recherche.htm). Эта база данных содержит также индексы организаций в перечне CCSD (сервер HAL; см. http://import.ccsd.cnrs.fr/doc/?consultLabs).

Основной массив информации (реляционная УРБД для хранения метаданных и собственно УРБД) был размещён на серверах высокого класса надёжности. Была также предусмотрена функция создания резервной копии информации на удалённом сервере для снижения риска потери информации. ПО серверов обеспечивает доступ к информации при помощи инструментария Apache, PHP, MySQL. Сайт системы имеет трёхъязычный интерфейс (французский, английский и китайский языки). На сайте введена в действие система авторизации доступа. Ряд операций доступен только зарегистрированным пользователям. В то же время любой посетитель сайта имеет свободный доступ к каталогу данных, инструментов и ресурсов, размещённых на сервере CRDO, а также к части относящихся к ним метаданных. В дальнейшем планируется также размещение в открытом доступе образцов аудио- и видеоматериалов УРБД, включая относящиеся к ним маркеры различных уровней (сегментная структура, просодическое оформление речи, данные о специфике артикуляции и т. п.).

Система встроенных запросов позволяет выбирать информацию из УРБД по параметрам, указанным в полях метаданных.

В дальнейшем CRDO планирует выполнить следующие работы:

— обеспечить возможность автоматического просмотра метаданных лингвистического характера в файле XML в формате OLAC для того, чтобы автоматизировать процедуру пополнения базы данных CRDO ссылками;

— ввести в действие систему текстовых информационных пространств Wiki для совместного редактирования веб-страниц в дополнение к метаданным;

— разработать процедуры ускоренного доступа к ресурсам УРБД;

— провести детальный анализ метаданных с целью унификации и стандартизации информационных структур УРБД, исключения ошибок и сбоев при поиске информации по запросам.

Литература

1. Автоматизированное рабочее место эксперта-фоноскописта. Электронная энциклопедия, версия V1.0: http://www.estra.ru

2. *Андрющенко В.М*. Концепция и архитектура Машинного фонда русского языка. М., 1989.

3. *Белолипецкий С.И., Буря А.Г*. Специализированные СУБД для поддержки речевых баз данных // Сетевой электронный научный журнал «Системотехника». № 2. 2004. М.: МГИЭМ, 2004.

4. *Богданов Д.С., Брухтий А.В., Кривнова О.Ф., Подрабинович А.Я., Строкин Г.С.* Технология формирования речевых баз данных // Сб. «Организационное управление и искусственный интеллект». М.: Эдиториал УРСС, 2003.

5. *Богданов Д.С., Кривнова О.Ф., Подрабинович А.Я., Фарсобина В.В*. База речевых фрагментов русского языка ISABASE // Сб. «Интеллектуальные технологии ввода и обработки информации». М.: Эдиториал УРСС, 1998.

6. *Богуславский И.М., Григорьев Н.В.* и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // Труды Международного семинара ДИАЛОГ-2000. М., 2000. Т. 2. С. 41–47.

7. Корпусная лингвистика в России / Сост. *Е.В. Рахилина* и *С.А. Шаров* // Спец. выпуск журнала НТИ. М., 2003. Сер. 2. № 6, 10.

8. *Кривнова О.Ф., Захаров Л.М., Строкин Г.С*. Речевые корпусы (опыт разработки и использование). М.: МГУ [http://www.dialog-21.ru/archive_article.asp].

9. *Кривнова О.Ф., Захаров Л.М., Строкин Г.С*. Речевые корпусы (опыт разработки и использование) // Сборник трудов Международного семинара Диалог'2001 по компьютерной лингвистике и её приложениям (в двух томах); Т.2. Прикладные проблемы. М., 2001.

10. *Кривнова О.Ф.* Области применения речевых корпусов и опыт их разработки // Сборник трудов XVIII сессии РАО. М.: ГЕОС, 2006.

11. *Леонтьева Н.Н.* Автоматическое понимание текстов: Системы, модели, ресурсы. М.:Академия/ Academia, 2006.

12. *Потапова Р.К., Линднер Г*. Особенности немецкого произношения. М.: Высшая школа, 1991. 319 с.

13. *Потапова Р.К.* Лингвистическое обеспечение Электронной Энциклопедии, предназначенной для экспертов-фоноскопистов (русский язык). М.: ЭСТРА, CDROM, 1998–1999.

14. *Потапова Р.К*. Новые информационные технологии и лингвистика. 4-е изд., сущ. доп. М.: Эдиториал УРСС, 2005. 368 с.

15. *Потапова Р.К*. Речь: коммуникация, информация, кибернетика. М.: Радио и Связь, 1997. 528 с.

16. *Потапова Р.К.*. Тайна современного кентавра. М.: Радио и связь, 1992.

17. *Рыков В.В*. Корпус текстов — новый тип словесного единства // Труды Международного семинара ДИАЛОГ-2003. Протвино, 2003.

18. *Сичинава Д.В*. К задаче создания корпусов русского языка в Интернете // НТИ. М., 2002. Сер. 2. № 12.

19. *Скрелин П.А., Щербаков П.П*. Требования к современной фонетической базе данных для фундаментальных и прикладных исследований // Технологии информационного общества — Интернет и современное общество: труды VI Всероссийской объединенной конференции. Санкт-Петербург, 3–6 ноября 2003 г. СПб.: Изд-во Филологического ф-та СПбГУ, 2003. С. 62–63.

20. *Шаров С.А*. Параметры описания текстов корпуса, а также Корпусная лингвистика в России // НТИ. М., 2003. Сер. 2. № 5–6.

21. *Arlazarov V.L., Bogdanov D.S. Krivnova O.F., Podrabinovitch A.Ya*. Creation of Russian Speech Databases: Design, Processing, Development Tools // International Conference SPECOM'2004. Proceedings. S-Pb. Russia, 2004. Pp: 650–656.

22. *Barlow M*. Corpora for Theory and Practice // 1JCL. Amsterdam, 1996. № 1.

23. *Bel B., Blache P*. Le Centre de Ressource pour la Description de l'Oral // Corpus. Les enjeux de l'annotation. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, 2006. Pp.13–18.

24. *Bertrand R., Blache P., Espesser R., Ferre G., Meunier C., Priego-Valverde B., Rauzy S.* Le CID — Corpus of Interactional Data: Protocoles, Conventions, Annotations // Corpus. Les enjeux de l'annotation. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, 2006. Pp.31–60.

25. *Bohmova A*. Automatic Procedures in Tectogrammatical Tagging. //The Prague Bulletin of Mathematical Linguistics. Prague, 2001. № 76. P. 23–34.

26. *Collier A., Pace y M., Renouf A*. Refining the Automatic Identification of Conceptual Relations in Large-scale Corpora. // Proceedings of the Sixth Workshop on Very Large Corpora. Montreal, 1998.

27. *Delais — Roussarie E., Post B., Portes C*. Annotation prosodique et typologia. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence. Vol.25, 2006. p. 61–95.

28. *Greenstette G., Segond F*. Multilingual Natural Language Processing // IJCL. 1997. V.2. — № 1.

29. *Hajicovd E., Pajas P., Vesela K*. Corpus Annotation on the Tectogrammatical Layer: Summarizing of the First Stages of Evaluations // The Prague Bulletin of Mathematical Linguistics. Prague, 2002. № 77. P. 5–18.

30. International Journal of Corpus Linguistics (IJCL) / Ed. W.Teubert. Amsterdam, 1996–2001.

31. *Kibkalo A.A., Lotkov M.M*. Choice of Phonetic Alphabet for Russian LVCSR System // Proceedings of the International Workshop «Speech and Computer» SPECOM' 2003. (Moscow, 27–29 October, 2003) Moscow: MSLU, 2003. P. 102–105.

32. *Kucova L., Hajic ova E*. Prague Dependency Treebank: Enrichment of the Underlying Syntactic Annotation by Coreferential Mark-Up // The Prague Bulletin of Mathematical Linguistics. Prague, 2004. № 81. P. 23–34.

33. *Lee Y.-J., Choi D.-L., Um Y., Lee K.-H., Kim Y.-I., Kim B.-W*. Speech Resources at SITEC in Korea // Proceedings of the 10th International Conference SPEECH and COMPUTER (SPECOM' 2005) (Patras, Greece, 17–19 October, 2005) Patras, Moscow: MSLU, 2005. P. 579–582.

34. *Loseva E., Potapova R.* Speech variability of vibrants: phonetic database for English and German // Proceedings of the 10th International Conference Speech and Computer SPECOM' 2005, Patras, Moscow: MSLU, 2005.

35. *Marcus M.P., Santorini B., Marcinkiewicz M.A*. Building a Large Annotated Corpus of English: The Penn Treebank // Computational Linguistics. 1993. Vol.19. № 2. P. 313–30.

36. *Potapova R.K, Potapov V.V.* Database of forensic phonetics knowledges (as applied to electronic encyclopaedia for Russian experts) // Proceedings of the International Conference of IAFP, York, UK, 1999. P. 6–7.

37. *Shaikevich A*. The Computer Fund of Russian Language // IJCL.-Amsterdam, 1997. V.2. № 1. P. 163–167.

38. *Teubert W*. Corpus Linguistics and Lexicography // IJCL. Philadelphia, 2001.

39. htp://www.mdi.ru/aspnews/body/03.12.2001_39303.html

40. http://cfrl.ru

41. http://conf.infosoc.ru/03-r2f14.html

42. http://www.auditech.ru

43. http://www.auditech.ru

44. http://www.mdi.ru/aspnews/body/03.12.2001_39303.html

Потапова Родмонга Кондратьевна
академик Международной академии информатизации,
доктор филол. наук, профессор.
заслуженный работник Высшей школы РФ,
зав. отделением прикладной лингвистики,
зав. кафедрой прикладной и экспериментальной лингвистики,
директор Центра фундаментального и прикладного речеведения
Московского государственного лингвистического университета.
Специалист в области романо-германского языкознания,
общей и прикладной фонетики, теоретической, прикладной,
экспериментальной и математической лингвистики.
Автор свыше 450 научных и научно- методических публикаций.