СТАТИСТИЧЕСКИЙ АНАЛИЗ ФОНЕТИЧЕСКОЙ СТРУКТУРЫ РЕЧЕВОГО КОРПУСА ДЛЯ СИСТЕМ РАСПОЗНАВАНИЯ И СИНТЕЗА РЕЧИ

Б. М. Лобанов, Л. И. Цирульник

Объединённый институт проблем информатики НАН Беларуси
Минск, Беларусь
E-mail: lobanov@newman.bas-net.by, liliya_tsirulnik@ssrlab.com

В работе рассмотрены вопросы выбора речевого корпуса на основе анализа его фонетической структуры, а также использования корпуса в системах распознавания и синтеза речи. Приведены основные результаты статистического анализа фонетической структуры исследуемого корпуса, показывающие степень его покрытия сегментами с различной степенью детализации качественного и количественного фонетического описания звуков.

Ключевые слова: речевой корпус, текстовый корпус, фонема, аллофон, синтез речи, распознавание речи.

ВВЕДЕНИЕ

Речевые технологии, использующие системы распознавания и синтеза речи, развиваются весьма бурно, поскольку они имеют ряд преимуществ перед типовыми средствами общения человека с машиной, такие как естественность, оперативность, освобождение рук и зрения пользователя, возможность управления в экстремальных ситуациях.

Существуют задачи в речевых технологиях, общие для распознавания и синтеза речи в рамках данного языка. Одной из наиболее важных является задача создания речевой БД, её сегментация и маркировка на фонетически значимые единицы речи. Полученные сегменты сохраняются в базе данных и служат для обучения акустических моделей в системе распознавания речи, а также для генерации голоса конкретного диктора в системе синтеза речи по тексту.

Удачное формирование текстового и речевого корпусов, наряду с лингвистически обоснованной классификацией и выбором базовых фонетических сегментов, во многом определяет надёжность распознавания, разборчивость и естественность синтезируемой речи. Сформированный корпус должен удовлетворять следующим требованиям [1]:

– корпус должен быть фонетически репрезентативным, т.е. в фонетической транскрипции текста должны встречаться все основные фонемы речи и их варианты;

– созданный корпус должен быть фонетически сбалансированным, то есть распределение частот встречаемости фонем и других фонетических единиц в корпусе должно быть близким к теоретическому;

– объём корпуса должен быть, по возможности, минимальным.

В данной работе в качестве текстовой БД предлагается использовать таблицы ГОСТ 16600-72, разработанные высококвалифицированным коллективом лингвистов для измерения фразовой разборчивости речи, передаваемой по каналам связи [2]. Предполагается, что созданный на этой основе речевой корпус будет достаточно полно удовлетворять перечисленным выше требованиям, т.к. в соответствии с целями разработчиков ГОСТа, таблицы должны включать минимально возможное количество фонетически репрезентативных и сбалансированных фраз. Целью настоящей работы является проверка этого предположения путём статистического анализа фонетической структуры речевого корпуса, построенного на базе таблиц ГОСТа, а также получение объективной информации о реальных статистических характеристиках фонетических сегментов различного уровня.

ОПИСАНИЕ ЭКСПЕРИМЕНТА

Для статистического анализа фонетической структуры корпуса выделяются сегменты с различной степенью детализации описания фонетического качества звуков (фонемы, позиционные аллофоны, позиционно-комбинаторные аллофоны) и их фонетического количества (монофоны, дифоны, слоги).

Для обозначения фонем используются следующие символы: ***A***- а, ***O*** - о, ***U***- у, ***E***- э, ***Y***- ы, ***I***- и, ***P*** - п, ***P'*** – п', ***T*** - т, ***T'*** - т', ***K*** - к, ***K'*** - к', ***B*** - б, ***B'*** - б', ***D*** - д, ***D'*** - д', ***G*** - г, ***G'*** - г', ***C*** - ц, ***Ch'*** - ч, ***F*** - ф, ***F'*** - ф', ***S*** - с, ***S'*** - с', ***Sh*** - ш, ***Sh'*** - щ, ***H*** - х , ***H'*** - х', ***V*** - в, ***V'*** - в', ***Z*** - з, ***Z'*** - з', ***Zh*** - ж, ***R*** - р, ***R'*** - р', ***M*** - м, ***M'*** - м', ***N*** - н, ***N'*** - н', ***L*** - л, ***L'*** - л', ***J'*** – й, где знак «'» показывает мягкость фонемы.

Аллофоны – оттенки фонем в речевом потоке – характеризуются позицией относительно полноударного гласного и фонемным окружением – предшествующей и последующей фонемами.

Аллофон обозначается именем фонемы и следующими за ним тремя целочисленными индексами: ***i, j, k*** [3]. Индекс ***i*** для гласной указывает её позицию по отношению к словесному ударению и может принимать одно из следующих значений: 0 – полноударный,1 – частично-ударный, 2 – первый предударный, 3 – заударный.

Индекс ***j*** указывает группу левого контекста, и принимает для гласных одно из следующих значений: 0 - синтагматическая пауза, 1 - твёрдые губные согласные, 2 - передне- и среднеязычные согласные, 3 - твёрдые заднеязычные согласные и гласные, 4 – мягкие согласные.

Индекс ***k*** указывает группу правого контекста, и для гласных означает следующее: 0 - синтагматическая пауза, 1 - твёрдые губные согласные, 2 - переднеязычные и заднеязычные твёрдые согласные и гласные, 3 - мягкие согласные.

Для обработки входного текста и статистического анализа его фонетической структуры используется разработанная ранее технология синтеза «речевых клонов» [4]. Процедура обработки входного текста (рис. 1) состоит из нескольких этапов. На первом этапе орфографический текст подвергается преобразованию буква-фонема (Б-Ф), происходит объединение фонем в дифонемы и фонослоги. На втором этапе полученная последовательность фонем подвергается преобразованию фонема – позиционный аллофон (Ф-ПА), полученные позиционные аллофоны объединяются в последовательности позиционных диаллофонов и позиционных аллослогов. Третий этап обработки текста включает преобразование позиционный аллофон – позиционно-

комбинаторный аллофон (ПА-ПКА), объединение аллофонов в диаллофоны и аллослоги. Последовательности данных, полученные на каждом этапе обработки текста (обозначенные на рис. 1 цифрами от 1 до 9), подаются на статистический анализатор, определяющий частоту встречаемости фонетических сегментов (дифференциальные распределения) и вычисляющий на этой основе степень покрытия корпуса различными элементами (интегральные распределения). Примеры полученных дифференциального и интегрального распределений для слоговых фонемных сегментов представлены на рис. 2.

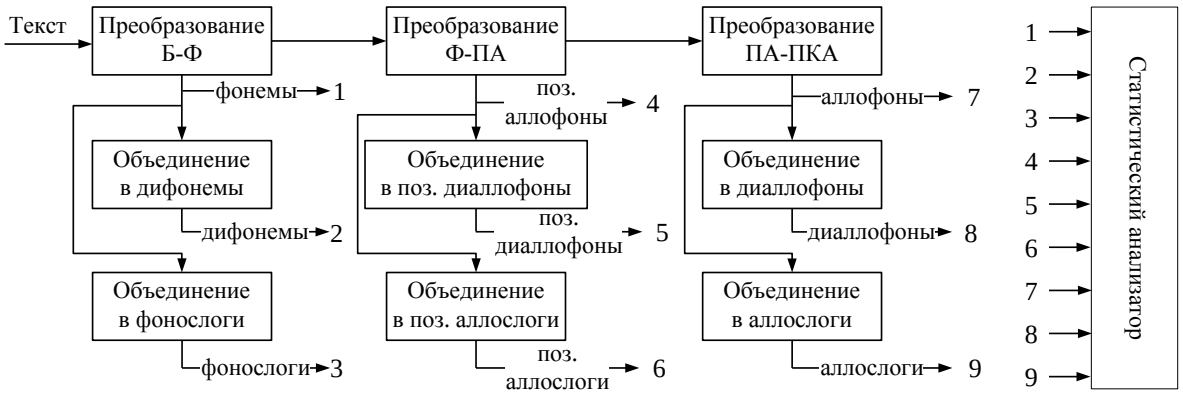
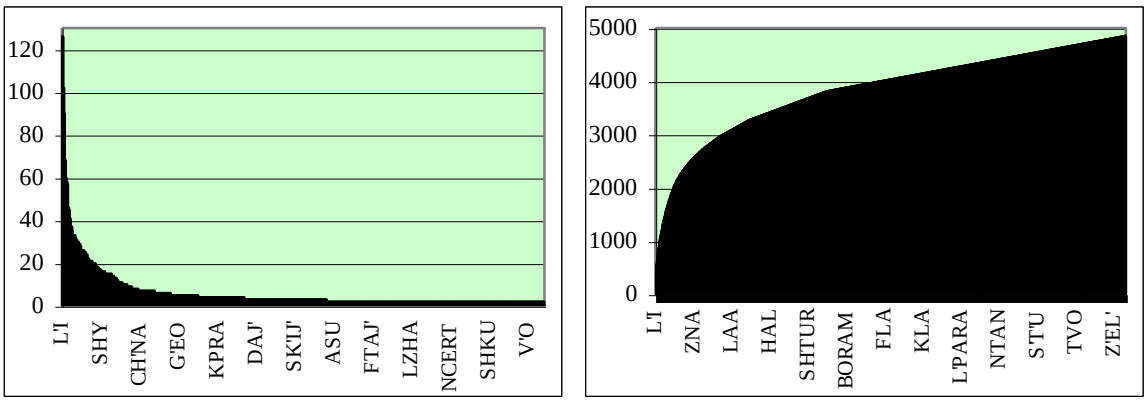


Рис.1. – Обработка текста и статистический анализ фонетической структуры корпуса



а) б)

Рис.2. –Дифференциальное (а) и интегральное (б) распределения слоговых фонемных сегментов

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

На рис. 3 приведены основные результаты статистического анализа фонетической структуры исследуемого речевого корпуса, показывающие степень его покрытия сегментами с различной степенью детализации качественного и количественного фонетического описания звуков. При этом по оси абсцисс отложено количество различных фонетических сегментов заданного типа- N_d, встретившихся в исследуемом тексте хотя бы один раз, а по оси ординат - процентное отношение общего количества сегментов заданного типа (различных и повторяющихся) к суммарному количе-

ству фонетических сегментов - N_s, встретившихся в тексте. Численные значения N_d и N_s для различных типов фонетических сегментов приведены в таблице 1.

Как видно из рис. 3, характер интегральных распределений для моносегментов различного фонетического качества (рис. 3 а,б,в) существенно отличаются от интегральных распределений для полисегментов различного фонетического количества (рис. 3 г,д,е). Для достижения 90 %-й степени покрытия текста требуемое количество различных сегментов уменьшается (от 69% до 43%) при увеличении степени детализации фонетического описания качества сегментов (фонемы, позиционные аллофоны, позиционно-комбинаторные аллофоны). В то же время, необходимое число различных сегментов увеличивается (от 43% до 84%) при увеличении фонетического количества (аллофоны, диаллофоны, аллослоги).

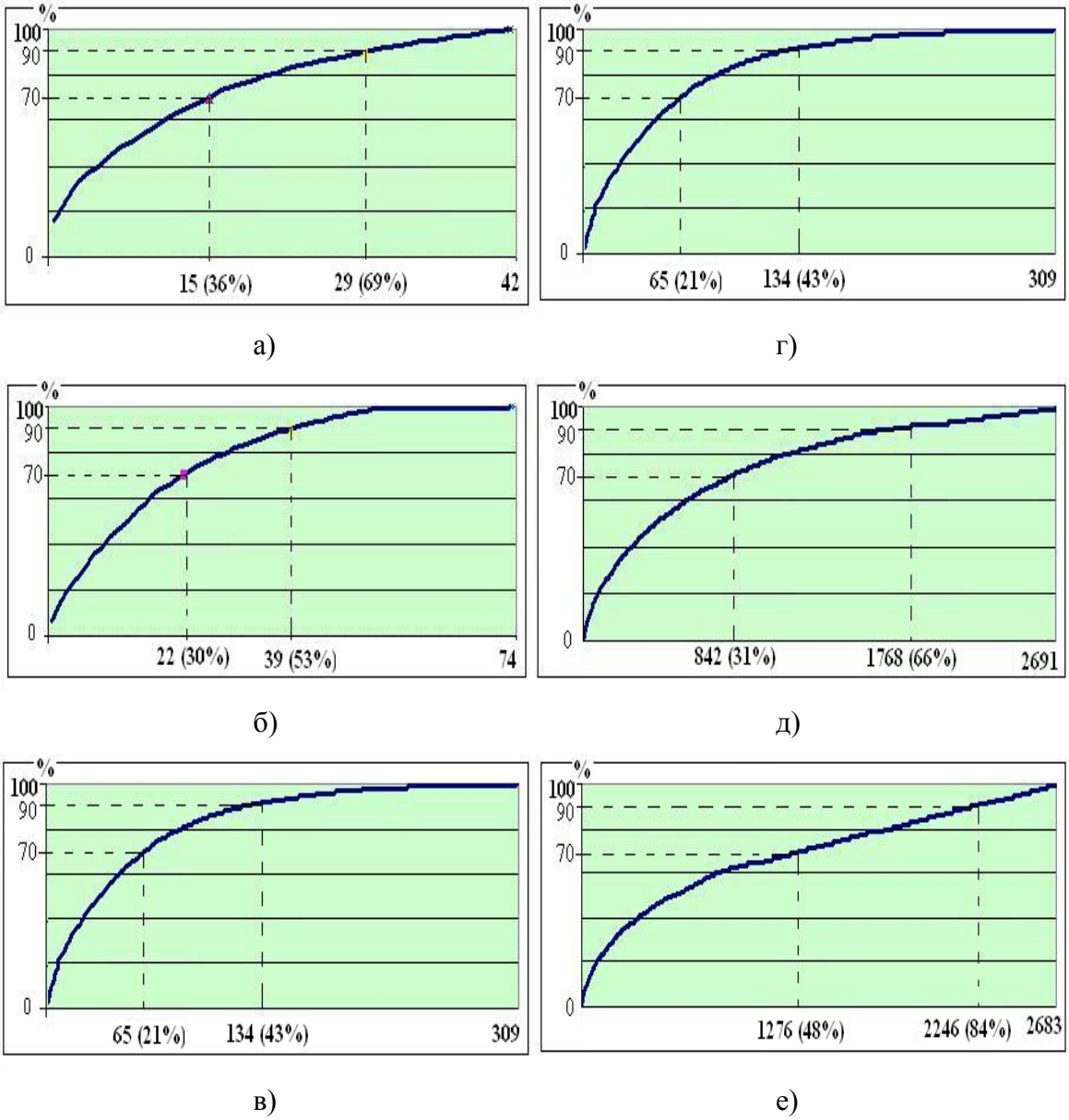


Рис.3. – Степень покрытия корпуса сегментами различного фонетического качества: а) фонемами, б) позиционными аллофонами, в) аллофонами; и различного фонетического количества: г) аллофонами, д) диаллофонами, е) аллослогами

Таблица 1

Численные значения различных фонетических сегментов и суммарное количество фонетических сегментов в корпусе

Тип сегментов	фонемы	поз. аллофоны	аллофоны	дифонемы	поз. диаллофоны	диаллофоны	фонослоги	поз. аллослоги	аллослоги
N_d	42	74	309	756	1326	2691	1638	2040	2683
N_s	10746	10746	10746	10247	10247	10247	4851	4851	4851

В таблице 2 приведены примеры десяти наиболее частотных фонетических сегментов различного типа, полученные в результате проведенного статистического анализа исследованных текстов.

Таблица 2

Десять наиболее частотных фонетических сегментов различного типа

Тип сегментов	1	2	3	4	5	6	7	8	9	10
фонемы	***A***	***I***	***E***	***R***	***T***	***N***	***O***	***L***	***U***	***K***
поз. аллофоны	A_3	A_2	R_0	I_3	A_0	T_0	N_0	O_0	L_0	K_0
аллофоны	T_{001}	R_{002}	S_{001}	N_{002}	P_{001}	L'_{002}	K_{002}	L_{002}	N'_{002}	R'_{002}
дифонемы	***RA***	***NA***	***L'I***	***PA***	***LA***	***KA***	***AL***	***AR***	***TA***	***N'I***
поз. диаллофоны	L'_0I_3	N_0A_3	R_0A_2	S_0T_0	L_0A_3	J'_0E_3	R_0A_3	K_0A_3	O_0T_0	P_0A_2
диаллофоны	S_{001} T_{001}	I_{343} J'_{012}	J'_{012} E_{342}	L'_{002} I_{342}	P_{001} R_{002}	R_{002} A_{223}	L'_{002} I_{343}	N'_{002} I_{343}	S'_{001} T'_{001}	P_{001} R'_{002}
фонослоги	***L'I***	***NA***	***PA***	***LA***	***ZA***	***MA***	***KA***	***VA***	***RA***	***TA***
поз. аллослоги	L'_0I_3	N_0A_3	L_0A_3	P_0A_2	P_0A_3	V_0A_3	N_0A_2	K'_0I_3	K_0A_2	R_0A_2
аллослоги	L'_{002} I_{342}	L'_{002} I_{343}	L_{002} A_{313}	L'_{002} I_{341}	P_{001} A_{212}	L_{002} A_{312}	N_{002} A_{322}	P_{001} A_{312}	N_{002} A_{323}	N_{002} A_{321}

ЗАКЛЮЧЕНИЕ

В настоящее время данные, полученные в результате проведенного статистического анализа исследованных текстов, активно используются в разрабатываемых системах синтеза «текст – речь» и системах распознавания «речь – текст» [5]. Для системы синтеза «текст – речь» в качестве звуковой БД отобрано свыше 6000 фонетических сегментов различного уровня, оптимальных с точки зрения получения синтезированной речи, близкой по качеству к натуральной. В системе распознавания «речь – текст» полученная информация активно используется для априорной оценки

информативности сегментов различного уровня, что позволяет существенно повысить общую надёжность распознавания речи.

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

ЛИТЕРАТУРА

1. Godfrey J., Zampolli A. Language Resources. Overview // Survey of the state of the art in human language technology. Chapter 12. Cambridge, 1996. http://cslu.cse.ogi.edu/HLTsurvey/ch12node3.html#SECTION121. Дата доступа: 1 Октября 2006 г.

2. ГОСТ 16600-72. Передача речи по трактам радиотелефонной связи. Москва, 1973.

3. Lobanov B.M., Tsirulnik. L.I. Development of multi-voice and multi-language TTS synthesizer (languages: Belarussian, Polish, Russian) // Proc. of the International Conference SPECOM'2006 (St. Petersburg 25-29 June 2006) / St. Petersburg: Anatoliya – p. 274-283.

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

5. Ронжин А.Л., Карпов А.А., Лобанов Б.М., Цирульник Л.И., Йокиш О. Фонетико-морфологическая разметка речевых корпусов для распознавания и синтеза русской речи // Информационно-управляющие системы. 2006. В печати.