ПРАВИЛА РАЗМЕТКИ РЕЧЕВОГО КОРПУСА НА ФОНЕТИЧЕСКИЕ СЕГМЕНТЫ И СТРАТЕГИЯ ВЫБОРА ЭЛЕМЕНТОВ КОМПИЛЯЦИИ ПРИ СИНТЕЗЕ РЕЧИ[1]

RULES OF SPEECH CORPUS SEGMENTATION INTO PHONETIC UNITS AND THE STRATEGY OF UNIT SELECTION IN SPEECH SYNTHESIS

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

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

Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

Рассмотрены варианты разметки речевого корпуса на внутрисловные и внутрисинтагменые фонетические сегменты: аллофоны, диаллофоны и аллослоги 3-х типов. Описаны алгоритмы разметки речевого корпуса на фонетические сегменты, приведены их статистические характеристики и стратегия выбора при синтезе речи.

Введение

Современные системы синтеза речи по тексту основаны на компиляции в непрерывный речевой сигнал фонетических сегментов размеченного речевого корпуса. При создании БД элементов компиляции существует несколько подходов, в соответствии с которыми могут быть сформированы базовые сегменты различной длины. При выборе сегментов той или иной длины разработчики систем синтеза речи используют, как правило, следующие критерии:

- объём работы, необходимый для создания речевого корпуса, последующей сегментации и маркировки;
- степень сохранения эффектов взаимодействия звуков, реализующиеся в естественном потоке речи;
- степень сохранения специфики межзвуковых переходов между выбранными элементами в естественном потоке речи.

При использовании звуковых единиц бóльшей длины в значительной степени сохраняется естественность эффектов взаимодействия звуков и характеристик межзвуковых переходов, но при этом резко возрастает количество звуковых единиц и, соответственно, объём работы для создания корпуса, его сегментации и маркировки. При использовании коротких речевых единиц меньше времени и усилий тратится на создание индивидуализированной речевой БД, но естественность проявления эффектов взаимодействия звуков и характеристик межзвуковых переходов могут быть представлены в недостаточной степени.

В данной работе предлагается компромиссный вариант разметки речевого корпуса на фонетические сегменты различного количественного и качественного состава и стратегия выбора элементов компиляции при синтезе речи по тексту.

1. Базовый набор сегментов для разметки текстов

В основу используемой нами классификации фонетических сегментов положено понятие аллофона – позиционного и комбинаторного оттенка фонемы. Как показал опыт синтеза речи по тексту, для русского языка, минимально-необходимый базовый набор аллофонов (мини-набор) должен включать 420 единиц (180 согласных и 240 гласных) [1]. Использование только базового набора аллофонов обеспечивает синтез вполне разборчивой речи по произвольному тексту, однако качество речи остаётся недостаточно высоким. Это объясняется тем, что реальное разнообразие оттенков фонем при их взаимодействии в потоке речи несоизмеримо большее, чем это обеспечивается используемым набором аллофонов. Кроме того, взаимовлияние соседних аллофонов в некоторых случаях может быть настолько сильным, что провести чёткую границу между ними зачастую просто невозможно. К таким случаям относятся, например, сочетания двух гласных аллофонов, а также некоторых сонорных согласных (таких, как /J/, /L/, /R/) и гласных. Существенное повышение качества и естественности речи может быть достигнуто, если в качестве элементов компиляции использовать не только аллофоны, но также и более протяжённые фонетические сегменты – мультифоны: диаллофоны, или ещё более протяжённые сегменты - аллослоги. Следует, однако, иметь в виду, что платой за достижение более высокого качества может стать резкое возрастание объёма БД элементов компиляции. Действительно, теоретический подсчёт количества возможных диаллофонов оценивается очень большим числом: $N_{da} = N^2_a = 420^2 = 176\ 400$. Далеко не все комбинации аллофонов возможны, но как показывает опыт, их количество в представительном речевом корпусе может достигать десятка тысяч.

Разметка корпуса осуществляется автоматически [2] с целью создания следующих групп фонетических

сегментов: аллофоны – $\{S_a\}$, диаллофоны – $\{S_{da}\}$, аллослоги – $\{S_{as}\}$. При создании БД элементов компиляции используются только наиболее частотные фонетические сегменты, извлекаемые из достаточно представительного корпуса естественной речи. Ниже, в разделе 3, рассмотрены характерные особенности статистических распределений для различных типов сегментов в используемом речевом корпусе [2].

Разметка корпуса на диаллофоны осуществляется посинтагменно, т.е. создаваемые диаллофонные сегменты могут находиться как внутри фонетического слова, так и на границе фонетических слов. Внутрисловное и межсловное различие существенно для диаллофонов типа ГГ и СГ (где Г обозначает гласный, С – согласный) [3]. Это различие в местоположении диаллофонов однозначно определяется позиционными индексами аллофонов.

Правила разметки речевого корпуса на слоговые комплексы учитывают фонетическую и артикуляционно-акустическую природу слога. Среди существующих определений понятия слог и способов слогоделения, описанных в литературе по фонетике [4-7], наиболее предпочтительным выглядит определение открытого СГ-слога, предложенное Л.В. Бондарко [7]. Это определение положено в основу разметки речевого корпуса на слоговые комплексы с некоторыми существенными уточнениями и дополнениями, вызванными объективными трудностями вычленения СГ-слога по следующим причинам:

- взаимная ассимиляция аллофонов в сочетаниях гласный-гласный, гласный-сонорный и в некоторых комбинациях сонорный-сонорный;
- редукция, вплоть до полного исчезновения, безударных гласных, находящихся между согласными.

В связи с этим при разметке создаются три типа слоговых комплексов, которые с точки зрения достижимой точности разметки являются трудно сегментируемыми (1-й тип), умеренной трудности сегментации (2-й тип), и относительно легко сегментируемыми (3-й тип).

При этом трудно сегментируемый слоговой комплекс ближе всего соответствует определению слога, данному в работе [7] и чаще всего будет иметь минимальную длительность. Слоговый комплекс умеренной трудности сегментации (который чаще всего будет иметь среднюю длительность) определяется в соответствии с контрастом смежных фонем по степени сонорности и в значительной степени соответствует определению слога, данному Л.В. Щербой и его последователями [8]. Границы легко сегментируемого слогового комплекса определяются с учётом указанных выше условий возможной полной редукции безударных гласных. Легко сегментируемый слоговый комплекс, как правило, будет иметь максимальную длительность.

2. Правила разметки на слоговые комплексы

Предлагаются следующие определения слоговых комплексов:

1. Слоговой комплекс 1-го типа определяется как открытый слог со следующими уточнениями:

- если за гласным, определяющим конец слога, находится гласный, он присоединяется к текущему слогу.

(1)[2] В слове "наивный", аллофонная запись "N_{002}, A_{223}, I_{021}, V_{013}, N_{002}, Y_{323}, J'_{010}", будут выделены следующие аллослоги (где границы аллослога помечаются символами "<", ">"): <$N_{002}A_{223}I_{021}$>, <$V_{013}N_{002}Y_{323}J'_{010}$>.

- если за гласным, определяющим конец слога, находится последовательность "*J'* – безударный гласный", вся последовательность присоединяется к текущему слогу.

(2) В слове "такая", аллофонная запись: "T_{001}, A_{222}, K_{002}, A_{033}, J'_{012}, A_{240}", будут выделены аллослоги <$T_{001}A_{222}$>, < $K_{002}A_{033}J'_{012}A_{240}$>.

- если слог состоит из одного гласного, он присоединяется к последующему слогу, формируя слоговой комплекс.

(3) В слове "аллофон", аллофонная запись: "A_{201}, L_{102}, A_{211}, F_{001}, O_{012}, N_{000}", будут выделены аллослоги <$A_{201}L_{102}A_{211}$>, <$F_{001}O_{012}N_{000}$>.

2. Для определения границ слогового комплекса 2-го типа выполняются описанные в п.1 правила и, кроме того, определены следующие дополнительные правила:

- если за гласным, определяющим конец слога, следует не менее двух согласных, первый из которых – сонант или *J', V, V'* (т.е. принадлежит множеству $\{J'_{ijk}, V_{ijk}, V'_{ijk}, R_{ijk}, R'_{ijk}, L_{ijk}, L'_{ijk}, N_{ijk}, N'_{ijk}, M_{ijk}, M'_{ijk}\}$), а последующий – нет, то граница определяется после первого из них.

(4) В слове "майка", аллофонная запись: "M_{002}, A_{013}, J'_{013}, K_{002}, A_{230}", будут выделены аллослоги <$M_{002}A_{013}J'_{013}$>, <$K_{002}A_{230}$>; в слове "жарко", аллофонная запись: "ZH_{002}, A_{022}, R_{003}, K_{002}, A_{230}", будут выделены аллослоги <$ZH_{002}A_{022}R_{003}$>, <$K_{002}A_{230}$>.

3. Для определения границ слогового комплекса 3-го типа выполняются описанные в п.1, 2 правила и, кроме того, определены следующие дополнительные правила:

- если за концом слога находится последовательность "сонант – безударный гласный", она присоединяется к текущему слогу.

(5) В слове "Ко́лоса", аллофонная запись: "K_{001}, O_{032}, L_{002}, A_{312}, S_{001}, A_{220}", будут выделены аллослоги

<$K_{001}O_{032}L_{002}A_{312}$>, <$S_{001}A_{220}$>.

- безударный слог, содержащий гласный второй степени редукции, присоединяется к предыдущему или последующему слогу, содержащему гласный меньшей степени редукции.

(7) В слове "фатализм", аллофонная запись: «F_{001}, A_{312}, T_{001}, A_{223}, L'_{002}, I_{042}, Z_{001}, M_{000}", будут выделены аллослоги <$F_{001}A_{312}T_{001}A_{223}$>, <$L'_{002}I_{042}Z_{001}M_{000}$>.

Разметка на слоговые комплексы каждого из перечисленных выше 3-х типов проводится двумя способами: пословно и посинтагменно. В первом случае разметка осуществляется независимо для каждого отдельного фонетического слова, входящего в синтагму. Во втором случае последовательность слов в синтагме рассматривается как единый речевой поток с учётом межсловных фонетико-акустический явлений, описанных в работе [3]. Очевидно, что поскольку на стыках слов могут встретиться любые сочетания фонем, невозможно создать речевой корпус разумного размера, в котором бы реализовались все сочетания. Целесообразно поэтому при использовании речевого корпуса воспользоваться обоими способами его разметки.

Таким образом, каждая речевая синтагма с учётом пословной и посинтагменной разметки и 3-х типов аллослогов размечается шестью различными способами. Пример разметки на различные виды слоговых комплексов синтагмы "Олимпийские чемпионы вернулись на родину" приведен в табл. 1, где границы аллослогов помечены значками "<", ">".

Вид разметки	**Тип слогового комплекса**	**Размеченная на слоговые комплексы синтагма** "Олимпийские чемпионы вернулись на родину"
Пословная	Трудно сегментируемый (тип 1)	<$A_{203}L'_{002}I_{243}$>, <$M'_{003}P'_{001}I_{043}$>, <$J'_{013}S_{001}K'_{001}I_{343}$>, <$J'_{012}E_{343}$>, <$CH'_{001}E_{343}$>, <$M'_{003}P'_{001}I_{241}O_{042}$>, <$N_{002}Y_{323}$>, <$V'_{012}E_{242}$>, <$R_{001}N_{002}U_{023}$>, <$L'_{002}I_{343}S'_{001}$>, <$N_{002}A_{222}$>, <$R_{002}O_{023}$>, <$D'_{002}I_{342}$>, <$N_{002}U_{320}$>
	Умеренной трудности сегментации (тип 2)	<$A_{203}L'_{002}I_{243}M'_{003}$>, <$P'_{001}I_{043}J'_{013}$>, <$S_{001}K'_{001}I_{343}J'_{012}E_{343}$>, <$CH'_{001}E_{343}M'_{003}$>, <$P'_{001}I_{241}O_{042}$>, <$N_{002}Y_{323}$>, <$V'_{012}E_{242}$>, <$R_{001}N_{002}U_{023}$>, <$L'_{002}I_{343}S'_{001}$>, <$N_{002}A_{222}$>, <$R_{002}O_{023}$>, <$D'_{002}I_{342}$>, < $N_{002}U_{320}$>
	Легко сегментируемый (тип 3)	<$A_{203}L'_{002}I_{243}M'_{003}$>, <$P'_{001}I_{043}J'_{013}$>, <$S_{001}K'_{001}I_{343}J'_{012}E_{343}$>, <$CH'_{001}E_{343}M'_{003}$>, <$P'_{001}I_{241}O_{042}N_{002}Y_{323}$>, <$V'_{012}E_{242}$>, <$R_{001}N_{002}U_{023}L'_{002}I_{343}S'_{001}$>, <$N_{002}A_{222}$>, <$R_{002}O_{023}$>, <$D'_{002}I_{342}N_{002}U_{320}$>
Посинтагменная	Трудно сегментируемый (тип 1)	<$A_{203}L'_{002}I_{243}$>, <$M'_{003}P'_{001}I_{043}$>, <$J'_{013}S_{001}K'_{001}I_{343}$>, <$J'_{012}E_{343}$>, <$CH'_{001}E_{343}$>, <$M'_{003}P'_{001}I_{241}O_{042}$>, <$N_{002}Y_{323}$>, <$V'_{012}E_{242}$>, <$R_{001}N_{002}U_{023}$>, <$L'_{002}I_{343}$>, <$S'_{001}N_{002}A_{222}$>, <$R_{002}O_{023}$>, <$D'_{002}I_{342}$>, <$N_{002}U_{320}$>
	Умеренной трудности сегментации (тип 2)	<$A_{203}L'_{002}I_{243}M'_{003}$>, <$P'_{001}I_{043}J'_{013}$>, <$S_{001}K'_{001}I_{343}J'_{012}E_{343}$>, <$CH'_{001}E_{343}M'_{003}$>, <$P'_{001}I_{241}O_{042}$>, <$N_{002}Y_{323}$>, <$V'_{012}E_{242}$>, <$R_{001}N_{002}U_{023}$>, <$L'_{002}I_{343}$>, <$S'_{001}N_{002}A_{222}$>, <$R_{002}O_{023}$>, <$D'_{002}I_{342}$>, <$N_{002}U_{320}$>
	Легко сегментируемый (тип 3)	<$A_{203}L'_{002}I_{243}M'_{003}$>, <$P'_{001}I_{043}J'_{013}$>, <$S_{001}K'_{001}I_{343}J'_{012}E_{343}$>, <$CH'_{001}E_{343}M'_{003}$>, <$P'_{001}I_{241}O_{042}N_{002}Y_{323}V'_{012}E_{242}$>, <$R_{001}N_{002}U_{023}L'_{002}I_{343}S'_{001}N_{002}A_{222}$>,<$R_{002}O_{023}$>, <$D'_{002}I_{342}N_{002}U_{320}$>

Табл 1. Разметка синтагмы на аллослоги

3. Статистический анализ фонетической структуры речевого корпуса

Статистический анализ корпуса, используемого для записей естественной речи и содержащего макси- и мини-тексты [2], проводился с целью выявления частоты встречаемости сегментов различного фонетического "качества" (фонемы, позиционные аллофоны, позиционно-комбинаторные аллофоны) и различного фонетического "количества" (аллофоны, диаллофоны, аллослоги). Процедура обработки текстового корпуса и получения статистических характеристик [9], показанная на рис. 1, состоит из нескольких этапов.

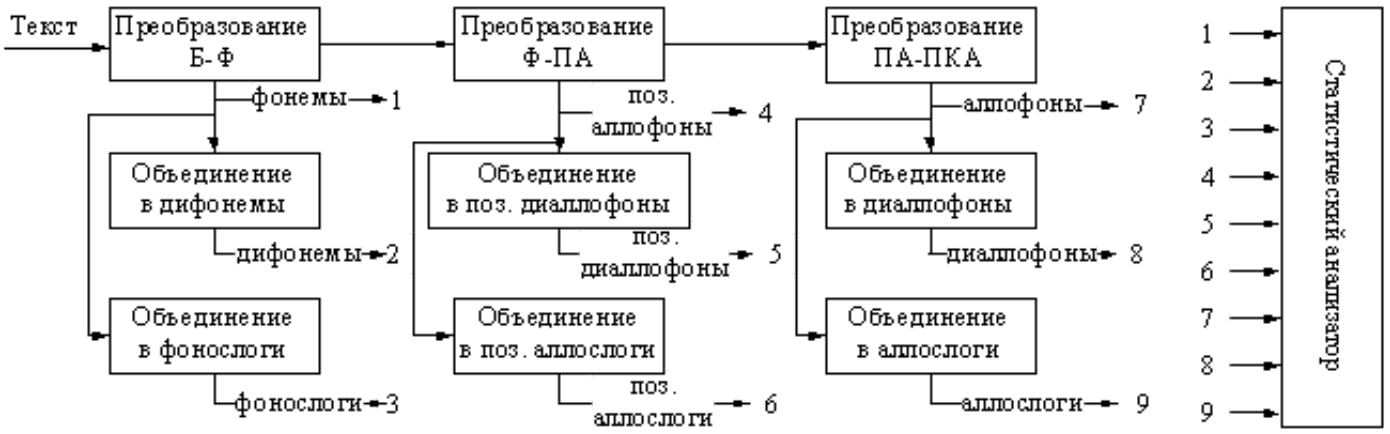


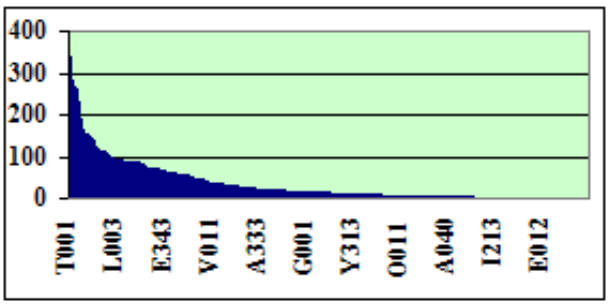
Рис. 1. Процедура обработки текста и статистического анализа фонетической структуры

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 текста различными элементами (интегральные распределения).

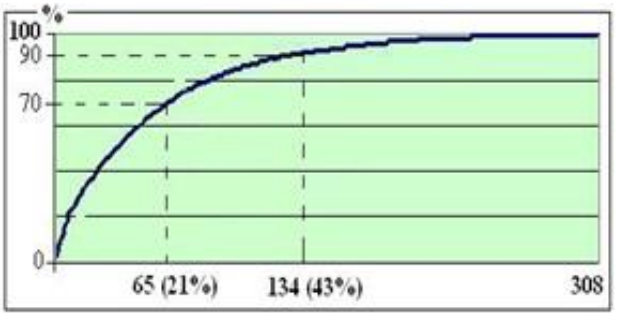
Дифференциальные и интегральные распределения частот встречаемости сегментов различного “фонетического количества” – аллофонов, диаллофонов, слогов – в макси-тексте представлены на рис. 2.

На графиках 2 а), б), в) показаны дифференциальные распределения для аллофонов, диаллофонов и аллослогов. По оси абсцисс расположены сегменты соответствующего типа в порядке уменьшения частоты их встречаемости в тексте, по оси ординат – количество сегментов указанного типа в тексте.

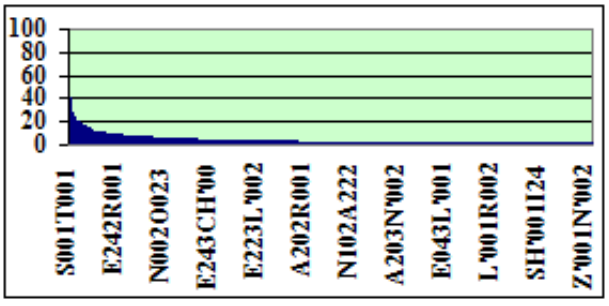
На графиках 2 г), д), е) показаны интегральные распределения аллофонов, диаллофонов и аллослогов в тексте. На каждом из графиков по оси абсцисс отложено количество различных фонетических сегментов заданного типа- N_d, а по оси ординат - процентное отношение общего количества сегментов заданного типа (различных и повторяющихся) к суммарному количеству фонетических сегментов - N_s, встретившихся в тексте.



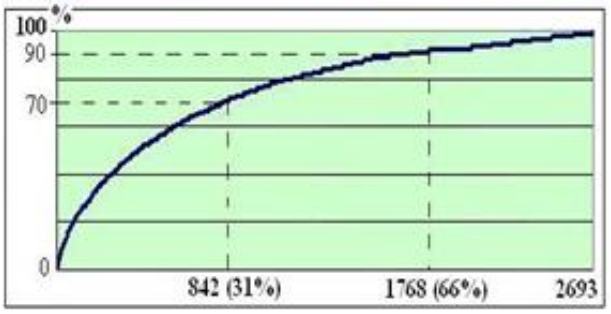
а)



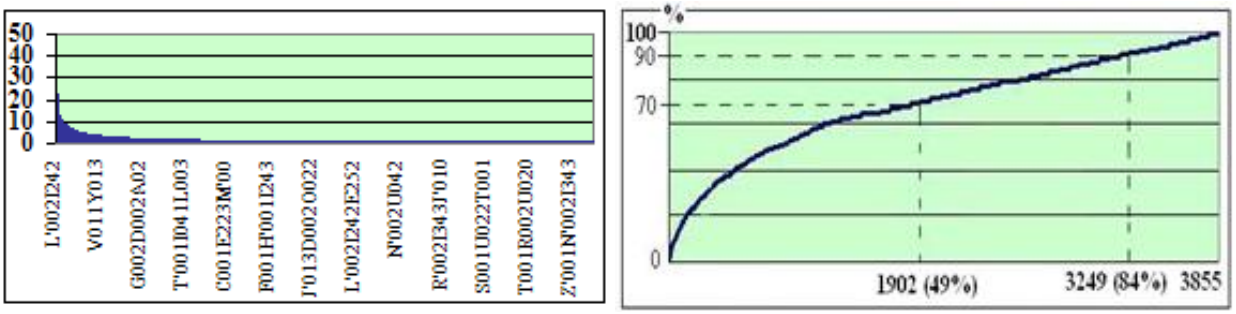
г)



б)



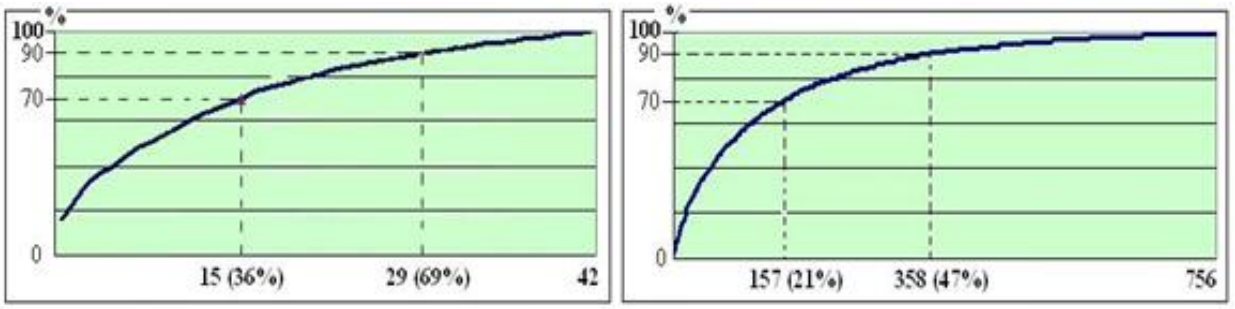
д)



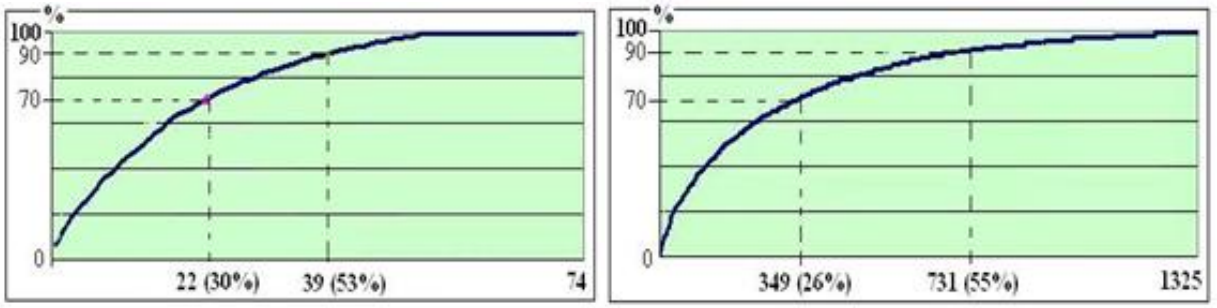
в) *е)*

Рис. 2. Дифференциальные (а, б, в) и интегральные (г, д, е) распределения в макси-тексте сегментов различного "фонетического количества": а, г – аллофонов; б, д – диаллофонов, в, е – аллослогов

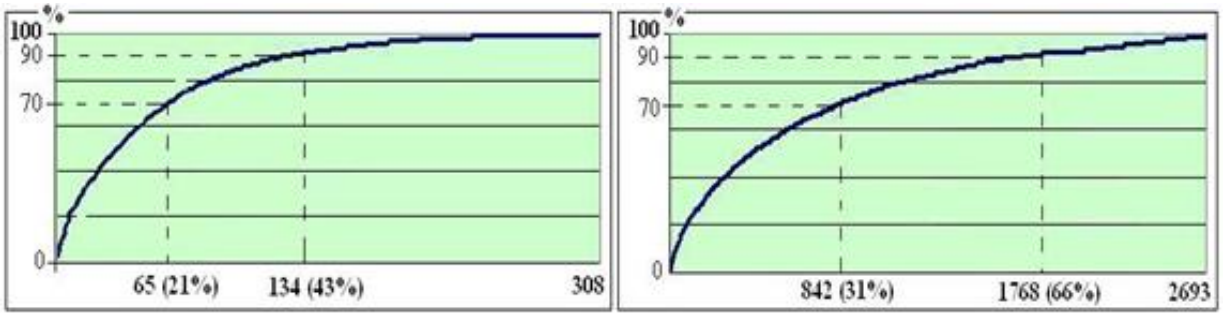
Степень покрытия макси-текста сегментами различного "фонетического качества": фоно-сегментами, позиционными алло-сегментами, позиционно-комбинаторными аллосегментами – представлена на рис. 3. По оси абсцисс отложено количество различных фонетических сегментов заданного типа- N_d, а по оси ординат - процентное отношение общего количества сегментов заданного типа к суммарному количеству фонетических сегментов - N_s.



а) *д)*



б) *е)*



в) *ж)*

Рис. 3. Степень покрытия текста сегментами различного "фонетического качества": а - фонемами, б - позиционными аллофонами, в - позиционно-комбинаторными аллофонами, г – дифонемами; д – позиционными диаллофонами, е – позиционно-комбинаторными диаллофонами

Как видно из графиков 2 – 3, характер интегральных распределений для полисегментов различного фонетического количества (рис. 2 г, д, е) существенно отличается от интегральных распределений для моносегментов различного фонетического качества (рис. 3). При увеличении показателя "фонетическое количество" требуемое для достижения 90 %-ой степени покрытия текста количество различных сегментов увеличивается от 43% до 84% (рис.

2 г, д, е). В то же время, увеличение степени детализации “фонетического качества” сегментов от фонем до аллофонов влечёт уменьшение от 69% до 43% необходимого числа различных сегментов (рис. 3 а, б, в), а увеличение степени детализации “фонетического качества” от дифонем до диаллофонов – увеличение от 47% до 66% необходимого для 90 %-ой степени покрытия текстов числа сегментов (рис. 3 г, д, е).

4. Стратегия выбора элементов компиляции при синтезе речи

На основе разметки речевого корпуса создаются соответствующие БД элементов компиляции: мини-набор аллофонов - $\{S_a\}$ и макси-набор мультифонов - $\{S_{as}, S_{da}\}$.

Аллофонная последовательность, формируемая в процессе синтеза речи по тексту, размечается на внутрисинтагменные и внутрисловные комплексы трёх типов. Затем осуществляется поиск в БД элементов компиляции полученных слоговых комплексов в соответствии со следующим приоритетом: внутрисинтагменные слоговые комплексы 3-го типа, внутрисловные слоговые комплексы 3-го типа, внутрисинтагменные слоговые комплексы 2-го типа, внутрисловные слоговые комплексы 2-го типа, внутрисинтагменные слоговые комплексы 1-го типа, внутрисловные слоговые комплексы 1-го типа. На каждом шаге выбора элементов компиляции из БД в случае, когда в БД не найден внутрисинтагменный слоговый комплекс 3-го типа, осуществляется последовательный поиск составляющих его слоговых комплексов других типов в соответствии с указанным выше приоритетом.

В случае, когда в БД не найден ни один из сформированных типов аллослогов, осуществляется поиск составляющих его диаллофонов. При этом всё множество диаллофонов разбивается в порядке уменьшения взаимовлияния соседних аллофонов и, как следствие, важности их вклада в качество синтезированной речи на 4 группы: ГГ, СГ, СС, ГС. Указанный порядок задаёт приоритет их выбора. В случае, когда необходимые диаллофоны отсутствуют в БД элементов компиляции, происходит выбор соответствующих аллофонов.

В результате указанной стратегии приоритетов элементы БД аллофонов, составляющие мини-набор, будут использоваться только в тех крайних случаях, когда необходимые для синтеза элементы верхних уровней – мультифоны – отсутствуют в имеющейся БД.

Заключение

Описанные выше правила разметки речевого корпуса на фонетические сегменты и стратегия выбора элементов компиляции реализованы в системе синтеза речи по тексту “МУЛЬТИФОН”. Их использование в системе позволило получить синтезированную речь с высокими показателями разборчивости и естественности. Образцы синтезированной речи будут продемонстрированы участникам конференции во время доклада.

Список литературы

1. Лобанов Б.М., Пьорковска Б., Рафалко Я., Цирульник Л.И., Шпилевский Э. Фонетико-акустическая база данных для многоязычного синтеза речи по тексту на славянских языках // “Компьютерная лингвистика и интеллектуальные технологии”: труды междунар. конф. Диалог’2006. М.: 2006. – С. 357–363.
2. Цирульник Л.И., Лобанов Б.М. Технология компьютерного клонирования и синтеза персональных характеристик речи диктора // “Компьютерная лингвистика и интеллектуальные технологии”: труды междунар. конф. Диалог’2007. М.: 2007. В печати.
3. Лобанов Б.М., Цирульник Л.И. Внутрисловные и межсловные правила обработки текста для полного и разговорного стилей речи // Функциональные стили звучащей речи: сб. науч. тр. М.: 2006. – С. 21–30.
4. Русская грамматика. М.: 1982. Т.1. С. 22–24.
5. Зиндер Л.Р. Общая фонетика. Л.: 1960.
6. Трахтеров А.Л. Основные вопросы теории слога и его определение // Вопросы языкознания. 1956. № 6. С 32–37.
7. Бондарко Л.В. Слоговая структура речи и дифференциальные признаки фонем (экспериментально-фонетическое исследование на материале русского языка) // автореф. дис. на соиск. учёной степени докт. филол. наук. Л.: 1969.
8. Грамматика русского языка. М.: 1952. Т.1. С. 71.
9. Лобанов Б.М., Цирульник Л.И. Статистический анализ фонетической структуры речевого корпуса для систем распознавания и синтеза речи // “Информационные системы и технологии”: материалы третьей междунар. конф. IST’2006. Мн.: 2006. Ч.2. С. 46–51.

---

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

[2] *В приводимых примерах аллофоны обозначаются именем фонемы и следующими за ней тремя индексами: i, j, k, где i указывает позицию фонемы, j – группу левого контекста, k – группу правого контекста*