Исследование статистических характеристик элементов фонетической структуры русской речи

Виталий Киселёв
vkiselov@newman.bas-net.by

В отличие от статистического анализа русскоязычных орфографических текстов, которому посвящено множество работ, исследования статистических характеристик элементов фонетической структуры русской речи практически отсутствуют. Между тем, такая информация представляет исключительную ценность в плане совершенствования качественных показателей систем синтеза речи по тексту.

Данная работа посвящена статистическому анализу фонетической структуры русской речи. Для успешного проведения эксперимента необходимо произвести предварительную обработку текстов. Для этого в орфографическом тексте расставляется автоматически ударение, расшифровываются числительные, идентифицируются аббревиатуры и сокращения. Затем преобразованный текст разбивается на акцентные группы, которые объединяются в синтагмы. Каждая синтагма переходит по правилам транскрибирования в фонемный вид. Такая фонемная последовательность и подвергается статистическому анализу. Был обработан большой массив транскрибированных таким образом текстов (свыше 34000000 знаков). Анализу подвергались следующие последовательности фонем: CC, CVV, VV, CV.

Результаты статистического анализа фонемного текста позволили выделить наиболее эффективный набор акустического инвентаря, который использовался для создания мультисегментного (полифонного) синтезатора речи, что существенно увеличило естественность и натуральность синтезированного речевого сигнала.

Введение

Изучение фонетической структуры речи, пожалуй, одна из наиболее важных задач для систем синтеза речи. Перед исследователями встаёт проблема нахождения оптимального состава фонетических единиц, их классификации, возможности группировки. В зависимости от целей и задач системы синтеза речи, возникает вопрос о выборе оптимального состава акустического инвентаря. Для целей, естественности и натуральности моделируемой речи, разработчики имеют тенденцию к увеличению акустического инвентаря, за счёт детализации фонетической структуры. Смешанный акустический инвентарь, начиная от аллофонов и заканчивая мультисегментами (полифонами), значительно повышает качество синтезируемой речи. Для аппаратной реализации синтеза речи, задача сводить к минимизации акустических баз путём объединения фонетических структур. Подобные модификации, возможны лишь в том случае, когда известна статистическая характеристика элементов фонетической структуры языка, в данном случаем русского.

Предварительная обработка

Для проведения исследования была создана программная среда, позволяющая производить статистический анализ различных фонетических последовательностей. В качестве входных данных система принимает фонемный текст. Для получения фонемного текст необходимо произвести нормализацию орфографического текста, включающую несколько этапов, за тем по правилам транскрибирования перевести в фонемный вид.

Алгоритм нормализации текста:

1. На первом этапе образуются орфографические синтагмы. Орфографической синтагмой считается последовательность слов объединенных по правилам:

- от начала текста до первого знака препинания;
- от знака препинания до знака препинания;
- от знака препинания до конца текста;

  Список знаков препинания {‘,’ ‘.’ ‘:’ ‘;’ ‘-‘ ‘?’ ‘!’}

2. Расшифровка сокращений, аббревиатур и числительных по определённым правилам.
3. Автоматическое расстановка ударения. В реализованной системе ударение расставляются посредством поиска необходимого слова в базе ударения. Если в базе присутствует слово, и оно не принадлежит к списку служебных слов, а именно союзам, предлогам и частицам тогда ему присваивается полное ударение (маркировка индексом 0). Если слово принадлежит списку служебных слов тогда ему присваивается частичное ударение (маркировка индексом 5). Если слово не нашлось, тогда все гласные маркируются как частично ударные (индекс 5).
4. Объединение слов в акцентные группы происходит посредством объединения слов с полным ударением со словами имеющие частичным ударением (служебные слова). Если слово принадлежит списку предлогов или списку союзов, тогда это слово присоединяется к последующему слову. Если слово принадлежит списку частиц (кроме частицы не), то оно присоединяется к предшествующему слову.
5. Формирование микро-синтагм. Формирование микро-синтагмы зависит от количества полных ударений в орфографической синтагмы. Если количество полных ударений меньше или равно четырём орфографическая синтагма остаётся без изменений. Если количество полных ударений больше четырёх - необходимо разбить синтагму на микро-синтагмы. Первым информативным маркером для членения является союзы и, или. Если данные союзы присутствуют в синтагме, тогда перед ними синтагма разделяется. После этого определяется количество полных ударений в каждых из синтагм. Если в какой-либо синтагме количество полных ударений превышает четырёх необходимо расчленить синтагму. Для этого, вначале расставляются маркеры, где слова не могут быть разделены. К классу таких слов относятся прилагательные. Тип слова определяется через базу ударения. Если в базе ударения отсутствует информация о типе слова, тогда прилагательное определяется с большой степенью вероятности по окончанию (ая, ее, его, ей, ему, ею, ие, ий, ими, их, ою, ого, ое, ой, ому, ою, ую, ый, ые, ым, ыми, ых, ою, яя). После того как расставлены маркеры неделимых слов происходит формирование микро-синтагм. Конец первой синтагмы ставиться после третьего слова с полным ударением, если за ним не стоит маркер неделимости. В случае если маркер стоит конец синтагмы ставиться после второго слова, и так же проверяется на наличия после него маркера неделимости. Данный цикл происходит итерационно до тех пор, пока в синтагме не останется меньше четырёх полных ударений.

После обработки текста происходит автоматическое транскрибирование. При этом ударные гласные маркируются индексом 0, предударные индексом 1, заударные индексом 2. В служебных словах ударная гласная маркируется индексом 5, а порядок индексирования заударных и предударных, при их наличии остаётся, прежней.

Методика проведения эксперимента

Анализу подвергался орфографический текст объёмом более 34000000 знаков. Электронные версии текста брались из библиотеки www.lib.ru, или аналога на компакт-диске. Из всего текста художественная проза составляла 25%, научно-популярные тексты, публицистика - 24%, драматургия 27%, электронные версии газет и журналов – 24%.

После нормализации текста и автоматического транскрибирования мы имеет фонемную строку вида: Ph1,Ph2,Ph3....,Phi,Phi+1,....Phn. Данная фонемная строка является микро-синтагмой фонемного уровня. В качестве, так называемых, псевофонем, выделен разрыв между акцентными группами, графическое обозначение которой “_” и границы микро-синтагмы – “##”. Таким образом, мы имеем общий вид фонемой строки: ##,Ph1,Ph2,Ph3,_,Ph4,Ph5,...,_,.....,##. Такая фонемная последовательность и подвергалась статистическому анализу. Были выделены несколько уровней, таких как двойное сочетание согласных CC, тройное сочетание согласных CCC, различные типы сочетаний гласных VV и сочетание согласных - гласный CV. Необходимо отметить, в выборке участвовали только те микро-синтагмы, у которых нет слов с неизвестной позицией ударения, т.е. если в базе ударения слово не присутствовала, тогда данная микро-синтагма не подвергалась анализу.

Фонемная статистика

В данной статье рассмотрены анализы сочетания фонем, являющиеся, по мнению автора, наиболее интересными для исследователей, следующее последовательности: двух и трёх согласных фонем (CC, CCC); двух гласных фонем (VV); сочетания согласный и гласный (CV). Необходимо отметить, что статистический анализ фонетической структуры русского языка, был проведёт для более расширенного состава сочетаемости фонем, например таких как CVC, выявляющее частотное окружение гласности, CVV, CCV, CVVC и других. Для каждой группы строилась результирующая таблица, выражающая процентное соотношение появления сочетания в рамках анализируемой группы и несколько орфографических примеров. Поскольку такая информация достаточна объёмна, она не была включена в данную статью.

На предварительном этапе фонемной статистики каждой фонеме ставиться атрибут гласности (V) и согласности (C). Фонемная строка разбивается на последовательность сегментов, в зависимости от задаваемого типа анализа и атрибута фонемы. Например, если нас интересует уровень CV, тогда фонема с атрибутом C объединяется в сегмент со следующей фонемой если она с атрибутом V. Анализ происходит до конца фонемной последовательности. Затем подсчитываются количество каждого встретившего сегмента, и делиться на общее количество сегментов группы. Для получения процентного соотношения, данное число умножается на 100%.

Результаты анализа приведены на графиках.

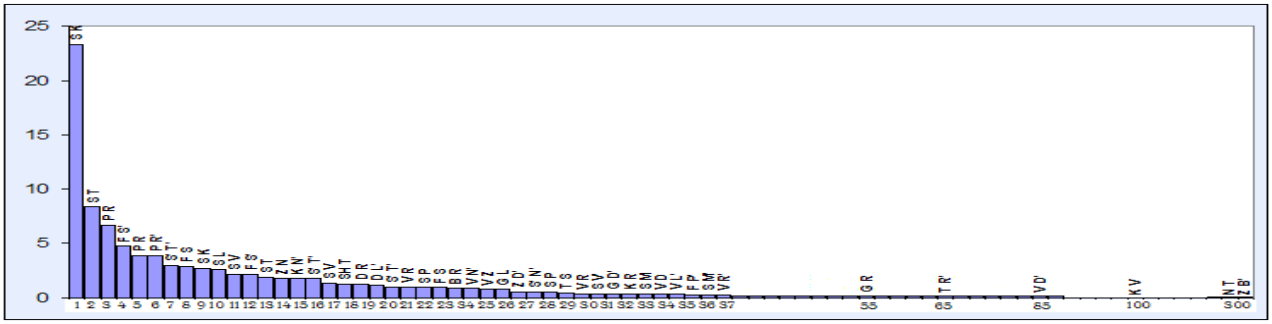

Рис. 1 Сочетание CC.

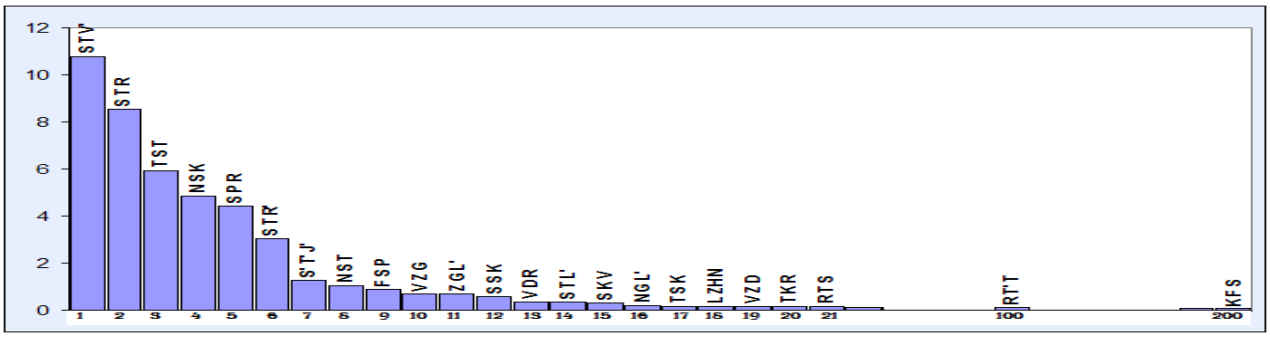

Рис. 2 Сочетание CCC.

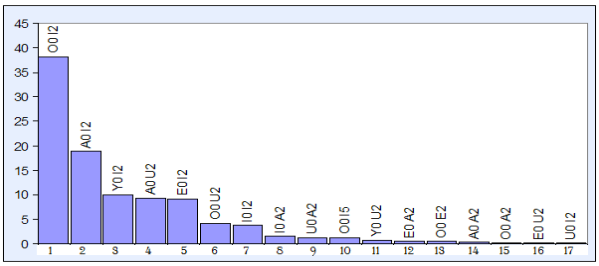

Рис. 3 Сочетание V0 V2.

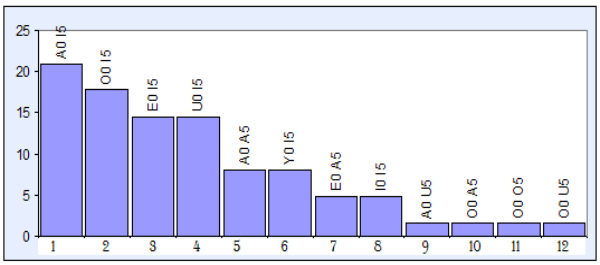

Рис. 4 Сочетание V0 V5.

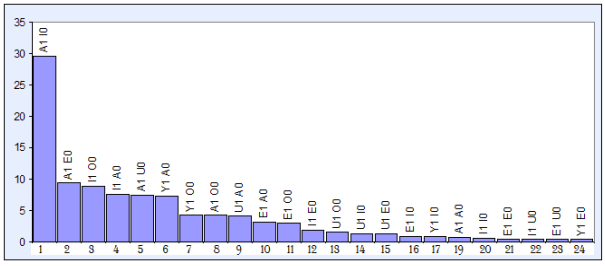

Рис. 5 Сочетание V1 V0.

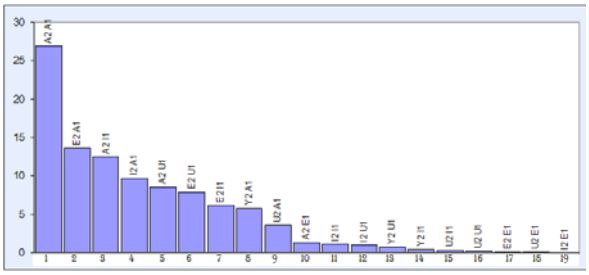

Рис. 10 Сочетание V2 V1

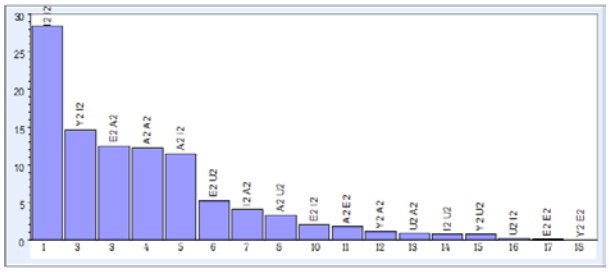

Рис.5 Сочетание V2 V2

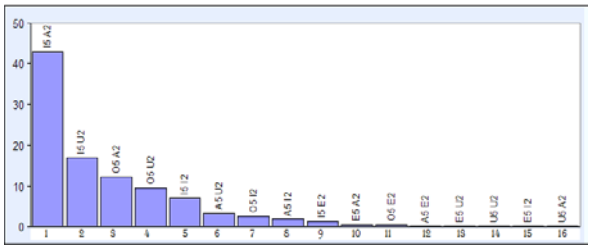

Рис.11 Сочетание V5 V2

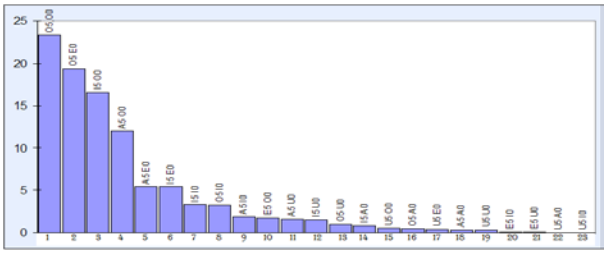

Рис.6 Сочетание V5 V0

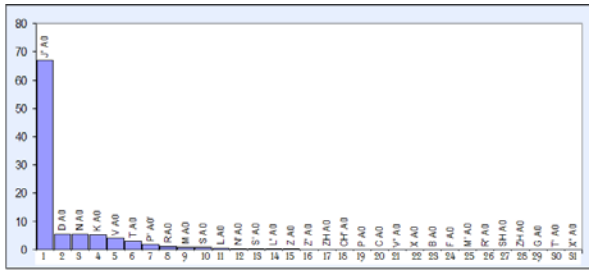

Рис.12 Сочетание C A0

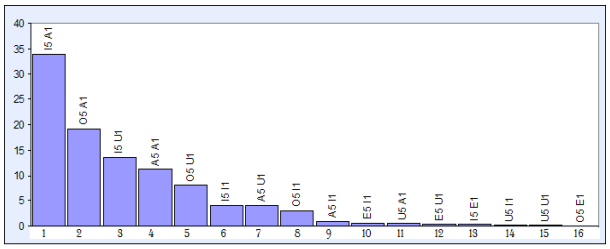

Рис.7. Сочетание V5 V1

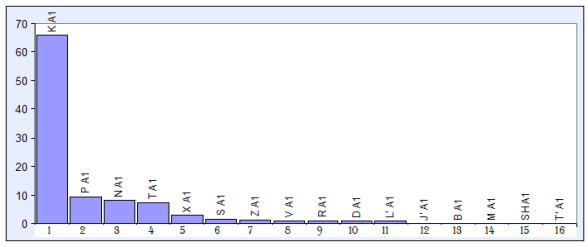

Рис.13 Сочетание C A1

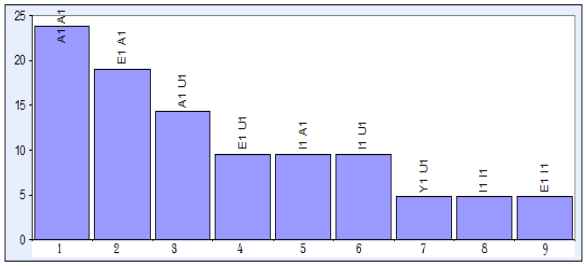

Рис. 8 Сочетание V1 V1.

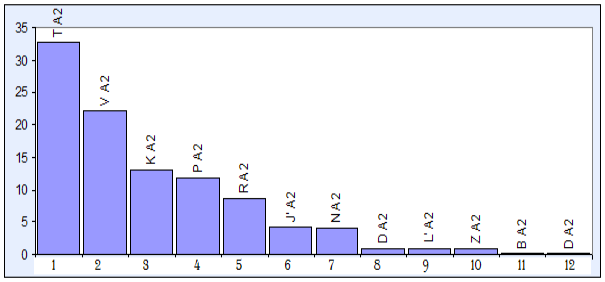

Рис.14 Сочетание C A2

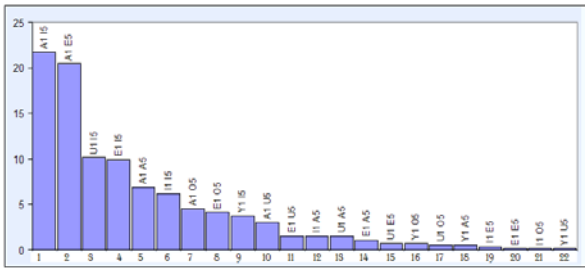

Рис.9 Сочетание V1 V5.

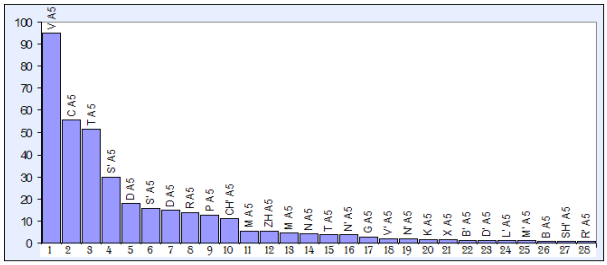

Рис.15 Сочетание C A5

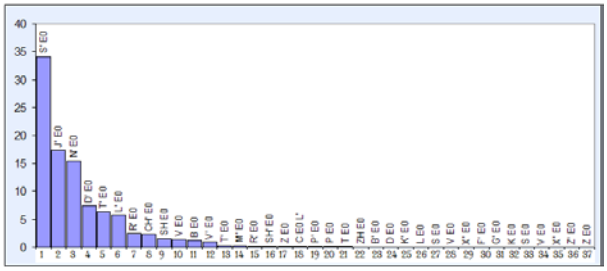

Рис.16 Сочетание С Е0

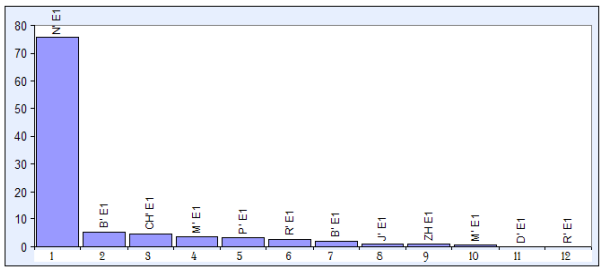

Рис.17 Сочетание С Е1

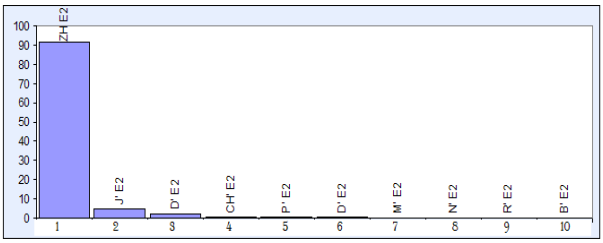

Рис.18 Сочетание С Е2

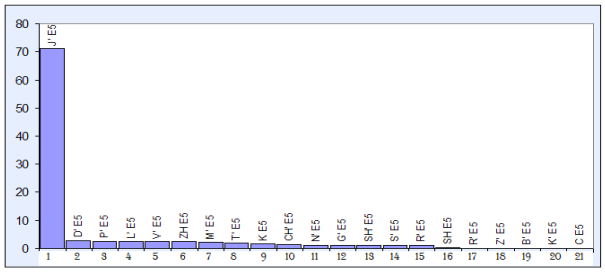

Рис.19 Сочетание С Е5

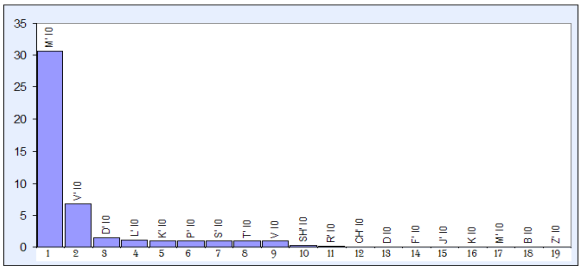

Рис.20 Сочетание С I0

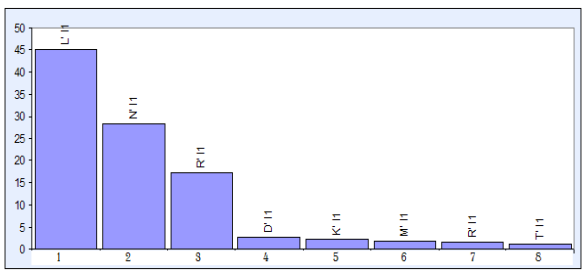

Рис.21 Сочетание С I1

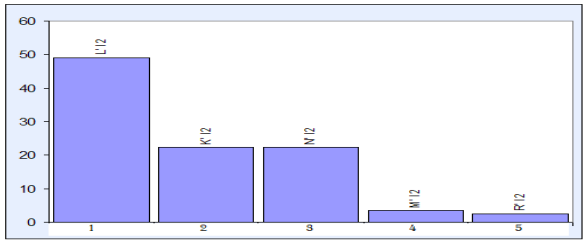

Рис.22 Сочетание С I2

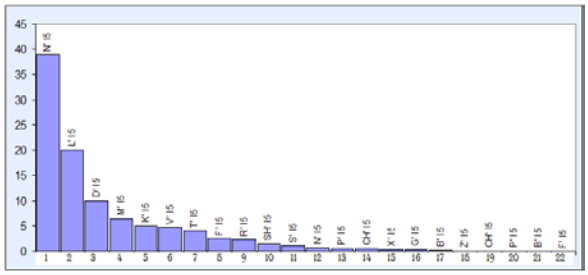

Рис.23 Сочетание С I5

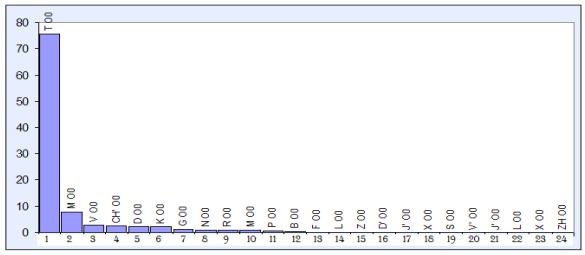

Рис.24 Сочетание С O0

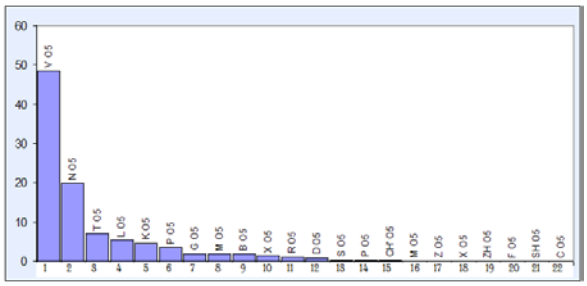

Рис.25 Сочетание С O5

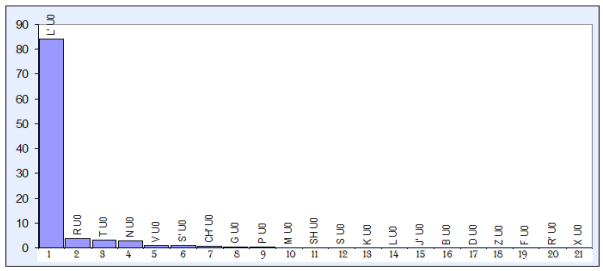

Рис.26 Сочетание С U0

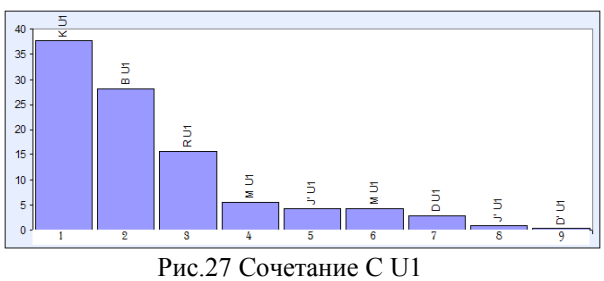

Рис.27 Сочетание С U1

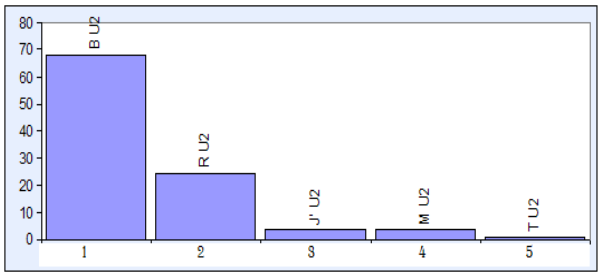

Рис.28 Сочетание С U2

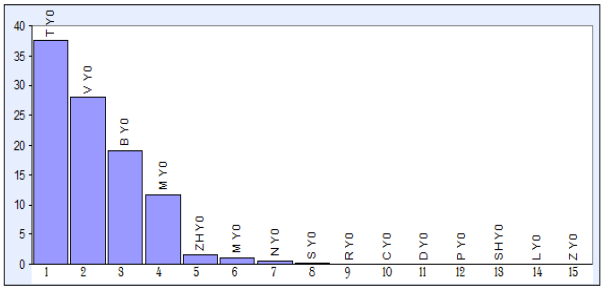

Рис.30 Сочетание С Y0

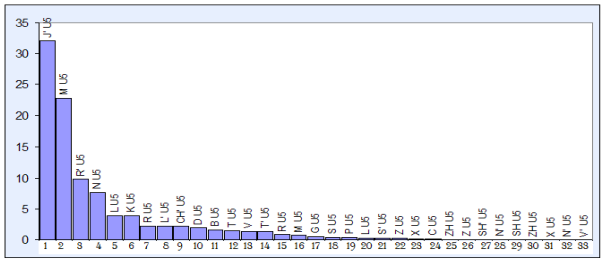

Рис.29 Сочетание С U5

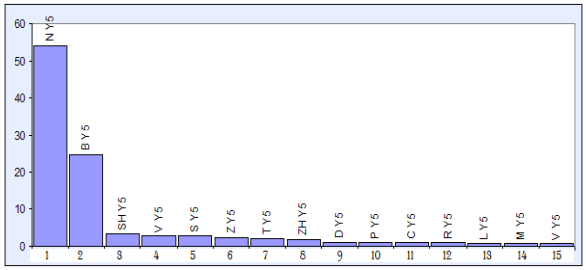

Рис.10 Сочетание С Y5

Заключение

Результаты статистического анализа фонемного текста позволили выделить наиболее эффективный набор акустического инвентаря, который использовался для создания мультисегментного (полифонного) синтезатора речи, что существенно увеличило естественность и натуральность синтезированного речевого сигнала.

Так же созданный механизм статистического анализа фонетической структуры позволит получить аналогичные данные и для других синтезируемых языков, разрабатываемых в лаборатории синтеза и распознавания речи.

Литература

1. Киселёв В.В., Лобанов Б.М., Левковская Т.В., Хейдоров И.Э. "Синтезатор персонализированной речи по тексту “ЛобаноФон-2000” //Тр. Международной конференции, посвящённой 100-летию российской экспериментальной фонетики.

Ст.-Петербург, 2001, С.101-104.

2. Киселёв В.В. "Аллофонный синтез русской речи по орфографическому тексту" // Сборник научных трудов. "Автоматическое распознавание и синтез речи", Минск, ИТК НАН Беларуси, 2000, С 155-163

3. Б.В. Сухотин. “Оптимизационные методы исследования языка” Москва 1976.