

06, июнь 2016

УДК 681.531.2

Алгоритм и методы распознавания речи

*Алборова Ж.В., студент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Робототехнические системы и мехатроника»*

*Научный руководитель: Рубцов В.И., к.т.н., доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Робототехнические системы и мехатроника»
kafsm7@sm.bmstu.ru*

Введение

Задача распознавания речи на сегодняшний день является актуальной проблемой. Большинство современных методов, используемых для ее решения, требуют больших вычислительных ресурсов, объем которых часто бывает ограничен. Невозможность широкого применения многих алгоритмов сегодня, например, в мобильных устройствах заставляет исследователей искать более эффективные методы.

Постановка задачи

Задачей данной работы является описание алгоритма и анализ методов распознавания речи, выявление недостатков каждого из них. Разработка программы по распознаванию речи и проведение эксперимента.

Общий алгоритм распознавания связной речи:

- Исходный сигнал
- Начальная фильтрация и усиление полезного сигнала
- Выделение отдельных слов
- Распознавание слова
- Распознавание речи
- Реакция на распознанный сигнал

Структурная система модуля распознавания изолированных слов приведена на рис. 1.

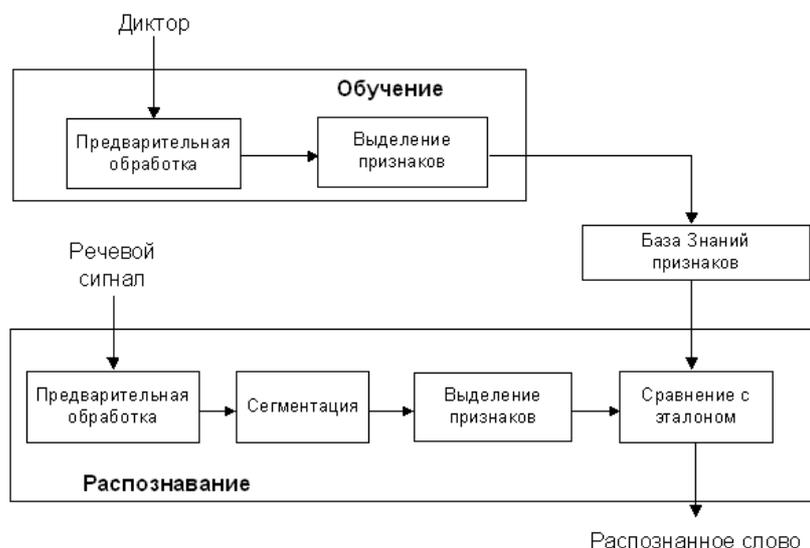


Рис. 1. Общая схема системы распознавания речи

Этапы распознавания:

1. Обработка речи начинается с оценки качества речевого сигнала. На этом этапе определяется уровень помех и искажений.

2. Результат оценки поступает в модуль акустической адаптации, который управляет модулем расчета параметров речи, необходимых для распознавания.

3. В сигнале выделяются участки, содержащие речь, и происходит оценка параметров речи. Происходит выделение фонетических и просодических вероятностных характеристик для синтаксического, семантического и прагматического анализа. (Оценка информации о части речи, форме слова и статистические связи между словами.)

4. Далее параметры речи поступают в основной блок системы распознавания — декодер. Это компонент, который сопоставляет входной речевой поток с информацией, хранящейся в акустических и языковых моделях, и определяет наиболее вероятную последовательность слов, которая и является конечным результатом распознавания.

Выделяют несколько основных способов распознавания речи:

1. Распознавание отдельных команд – раздельное произнесение и последующее распознавание слова или словосочетания из небольшого заранее заданного словаря. Точность распознавания ограничена объемом заданного словаря

2. Распознавание по грамматике – распознавание фраз, соответствующих определенным правилам. Для задания грамматик используются стандартные XML-языки, обмен данными между системой распознавания и приложением осуществляется по протоколу MRCP.

3. Поиск ключевых слов в потоке слитной речи – распознавание отдельных участков речи. Речь может быть как спонтанной, так и соответствующей определённым правилам. Произнесенная речь не полностью преобразуется в текст - в ней автоматически находятся те участки, которые содержат заданные слова или словосочетания.

4. Распознавание слитной речи на большом словаре – все, что сказано, дословно преобразуется в текст. Достоверность распознавания достаточно высока.

5. Распознавание речи с помощью нейронных систем. На базе нейронных сетей можно создавать обучаемые и самообучающиеся системы, что является важной предпосылкой для их применения в системах распознавания (и синтеза) речи.

а) Представление речи в виде набора числовых параметров. После выделения информативных признаков речевого сигнала можно представить эти признаки в виде некоторого набора числовых параметров (т.е. в виде вектора в некотором числовом пространстве). Далее задача распознавания примитивов речи сводится к их классификации при помощи обучаемой нейронной сети.

б) Нейронные ансамбли. В качестве модели нейронной сети, пригодной для распознавания речи и обучаемой без учителя можно выбрать самоорганизующуюся карту признаков Кохонена. В ней для множества входных сигналов формируется нейронные ансамбли, представляющие эти сигналы. Этот алгоритм обладает способностью к статистическому усреднению, что позволяет решить проблему изменчивости речи.

в) Генетические алгоритмы. При использовании генетических алгоритмов создаются правила отбора, позволяющие определить, лучше или хуже справляется новая нейронная сеть с решением задачи. Кроме того, определяются правила модификации нейронной сети. Изменяя достаточно долго архитектуру нейронной сети и отбирая те архитектуры, которые позволяют решить задачу наилучшим образом, рано или поздно можно получить верное решение задачи. [2]

Системы распознавания речи классифицируются:

- по размеру словаря (ограниченный набор слов, словарь большого размера);
- по зависимости от диктора (дикторозависимые и дикторонезависимые системы);
- по типу речи (слитная или отдельная речь);
- по назначению (системы диктовки, командные системы);
- по используемому алгоритму (нейронные сети, скрытые Марковские модели, динамическое программирование);
- по типу структурной единицы (фразы, слова, фонемы, дифоны, аллофоны);

- по принципу выделения структурных единиц (распознавание по шаблону, выделение лексических элементов).

Для систем автоматического распознавания речи, помехозащищённость обеспечивается, прежде всего, использованием двух механизмов:

- Использование нескольких, параллельно работающих, способов выделения одних и тех же элементов речевого сигнала на базе анализа акустического сигнала;
- Параллельное независимое использование сегментного (фонемного) и целостного восприятия слов в потоке речи.

Методы и алгоритмы распознавания речи

Сегодня системы распознавания речи строятся на основе принципов признания форм распознавания. Методы и алгоритмы, которые использовались до сих пор, могут быть разделены на следующие большие классы:

Классификация методов распознавания речи на основе сравнения с эталоном.

1. Динамическое программирование — временные динамические алгоритмы (Dynamic Time Warping).

Контекстно-зависимая классификация. При её реализации из потока речи выделяются отдельные лексические элементы — фонемы и аллофоны, которые затем объединяются в слоги и морфемы.

2. Скрытые Марковские модели (Hidden Markov Model);
3. Нейронные сети (Neural networks).

1. Использование DWT алгоритма в распознавании речи

Определение слов с использованием алгоритма DWT

Определение слова может осуществляться путем сравнения числовых форм сигналов или путем сравнения спектрограммы сигналов. Процесс сравнения в обоих случаях должен компенсировать различные длины последовательности и нелинейный характер звука. DWT алгоритму удастся разобрать эти проблемы путем нахождения деформации, соответствующей оптимальному расстоянию между двумя рядами различной длины.[4]

Существуют 2 особенности применения алгоритма:

1. Прямое сравнение числовых форм сигналов. В этом случае, для каждой числовой последовательности создается новая последовательность, размеры которой значительно меньше. Числовая последовательность может иметь несколько тысяч числовых значений, в то время как подпоследовательность может иметь несколько сотен значений.

Уменьшение количества числовых значений может быть выполнено путем их удаления между угловыми точками. Этот процесс сокращения длины числовой последовательности не должен изменять своего представления. Несомненно, процесс приводит к уменьшению точности распознавания. Однако, принимая во внимание увеличение скорости, точность, по сути, повышается за счет увеличения слов в словаре.

2. Представление сигналов спектрограмм и применение алгоритма DTW для сравнения двух спектрограмм. Метод заключается в разделении цифрового сигнала на некоторое количество интервалов, которые будут перекрываться. Для каждого импульса, интервалы действительных чисел (звуковых частот), будет рассчитывать Быстрым преобразованием Фурье, и будет храниться в матрице звуковой спектрограммы. Параметры будут одинаковыми для всех вычислительных операций: длин импульса, длины преобразования Фурье, длины перекрытия для двух последовательных импульсов. Преобразование Фурье является симметрично связанным с центром, а комплексные числа с одной стороны связаны с числами с другой стороны.

В связи с этим, только значения из первой части симметрии можно сохранить, таким образом, спектрограмма будет представлять матрицу комплексных чисел, количество линий в такой матрице является равной половине длины преобразования Фурье, а количество столбцов будет определяться в зависимости от длины звука. DTW будет применяться на матрице вещественных чисел в результате сопряжения спектрограммы значений, такая матрица называется матрицей энергии.

2. Применение скрытых Марковских моделей для распознавания речи

Скрытой Марковской моделью (СММ) называется модель состоящая из N состояний, в каждом из которых некоторая система может принимать одно из M значений какого-либо параметра. Вероятности переходов между состояниями задается матрицей вероятностей $A=\{a_{ij}\}$, где a_{ij} – вероятность перехода из i -го в j -е состояние. Вероятности выпадения каждого из M значений параметра в каждом из N состояний задается вектором $V=\{b_j(k)\}$, где $b_j(k)$ – вероятность выпадения k -го значения параметра в j -м состоянии. Вероятность наступления начального состояния задается вектором $\pi=\{\pi_i\}$, где π_i – вероятность того, что в начальный момент система окажется в i -м состоянии.

Таким образом, скрытой Марковской моделью называется тройка $\lambda=\{A,V,\pi\}$. Использование скрытых Марковских моделей для распознавания речи основано на двух приближениях:

1) Речь может быть разбита на фрагменты, соответствующие состояниям в СММ, параметры речи в пределах каждого фрагмента считаются постоянными.

2) Вероятность каждого фрагмента зависит только от текущего состояния системы и не зависит от предыдущих состояний.

Модель называется «скрытой», так как нас, как правило, не интересует конкретная последовательность состояний, в которой пребывает система. Мы либо подаем на вход системы последовательности типа $O=\{o_1,o_2,\dots,o_i\}$ - где каждое o_i – значение параметра (одно из M), принимаемое в i -й момент времени, а на выходе ожидаем модель $\lambda=\{A,B,\pi\}$ с максимальной вероятностью генерирующую такую последовательность, - либо наоборот подаем на вход параметры модели и генерируем порождаемую ей последовательность. И в том и другом случае система выступает как “черный ящик”, в котором скрыты действительные состояния системы, а связанная с ней модель заслуживает названия скрытой.

Для осуществления распознавания на основе скрытых моделей Маркова необходимо построить кодовую книгу, содержащую множество эталонных наборов для характерных признаков речи (например, коэффициентов линейного предсказания, распределения энергии по частотам и т.д.). Для этого записываются эталонные речевые фрагменты, разбиваются на элементарные составляющие (отрезки речи, в течении которых можно считать параметры речевого сигнала постоянными) и для каждого из них вычисляются значения характерных признаков. Одной элементарной составляющей будет соответствовать один набор признаков из множества наборов признаков словаря.[1]

Фрагмент речи разбивается на отрезки, в течении которых параметры речи можно считать постоянными. Для каждого отрезка вычисляются характерные признаки и подбирается запись кодовой книги с наиболее подходящими характеристиками. Номера этих записей и образуют последовательность наблюдений $O=\{o_1,o_2,\dots,o_i\}$ для модели Маркова. Каждому слову словаря соответствует одна такая последовательность. Далее A – матрица вероятностей переходов из одного минимального отрезка речи (номера записи кодовой книги) в другой минимальный отрезок речи (номер записи кодовой книги). B – вероятности выпадения в каждом состоянии конкретного номера кодовой книги (рис. 2).

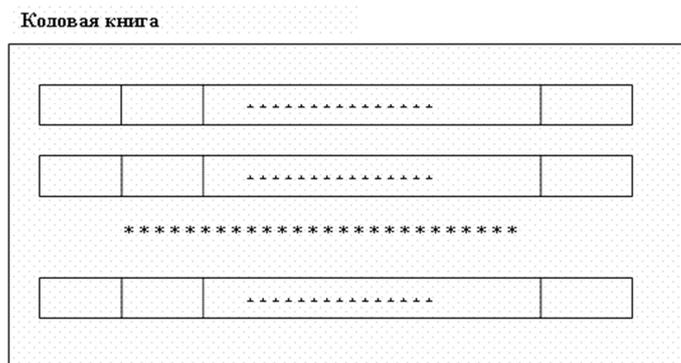


Рис. 2. Кодовая книга

На этапе настройки моделей Маркова мы применяем алгоритм Баума- Уэлча для имеющегося словаря и сопоставления каждому из его слов матрицы А и В.

При распознавании мы разбиваем речь на отрезки, для каждого вычисляем набор номеров кодовой страницы и применяем алгоритм прямого или обратного хода для вычисления вероятности соответствия данного звукового фрагмента определенному слову словаря. Если вероятность превышает некоторое пороговое значение – слово считается распознанным.

3. Применение нейронных сетей для распознавания речи

Искусственная нейронная сеть — это математическая модель, а также устройства параллельных вычислений, представляющие собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейронов). Как математическая модель искусственная нейронная сеть представляет собой частный случай методов распознавания образов или дискриминантного анализа. Пример нейросети изображен на рис.3.

Каждый процессор подобной сети имеет дело только с сигналами, которые он периодически получает, и сигналами, которые он периодически посылает другим процессорам. И тем не менее, будучи соединёнными в достаточно большую сеть с управляемым взаимодействием, такие локально простые процессоры вместе способны выполнять довольно сложные задачи.[6]

Понятие возникло при изучении процессов, протекающих в мозге при мышлении, и при попытке смоделировать эти процессы. Полученные модели называются искусственными нейронными сетями (ИНС).

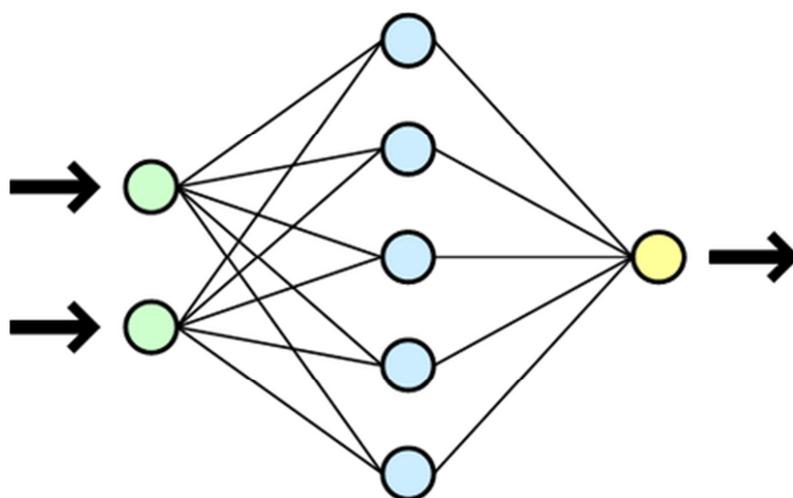


Рис. 3 Схема простой нейросети.

Зелёным обозначены входные элементы, жёлтым — выходной элемент

Нейронные сети не программируются в привычном смысле этого слова, они обучаются. Возможность обучения — одно из главных преимуществ нейронных сетей перед традиционными алгоритмами. Технически обучение заключается в нахождении коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными данными и выходными, а также выполнять обобщение. Это значит, что, в случае успешного обучения, сеть сможет вернуть верный результат на основании данных, которые отсутствовали в обучающей выборке.

Решение задачи распознавания с помощью нейронных сетей обладает значительным преимуществом перед алгоритмами, основанными на вычислении метрик — вычислительные затраты не зависят от количества слов в словаре. При увеличении длины словаря увеличивается лишь размер обучающей выборки, то есть нейронной сети требуется затратить больше времени на процесс обучения, но трудоемкость процесса распознавания не изменяется. Такая особенность позволяет оперировать с достаточно большим количеством слов в словарях. Недостатком нейросетевого подхода является отсутствие возможности добавления новых слов в словарь после окончания процесса обучения. Для разрешения этой проблемы может быть применена теория адаптивного резонанса. Нейронные сети, построенные в рамках теории адаптивного резонанса сохраняют пластичность при запоминании новых образов, и, в то же время, предотвращают модификацию старой памяти.[3]

Экспериментальная часть

Мною была написана программа в среде Matlab и проведены экспериментальные исследования алгоритма распознавания речевых сигналов

Стояла задача собрать данные по распознаванию двух слов: «Hello» и «Start». Каждое слово было произнесено пятью людьми по пять раз. Был установлен шумовой порог, т.к. шумы хоть и были незначительны, но все же могли повлиять на результаты.

Результаты статистических данных приведены в таблицах.

Алгоритм программы

В начале работы на экран выводится главное окно программы. После этого на динамик микрофона подается звуковое сообщение, за который отвечает модуль ввода речевого сигнала. Затем на главном окне пользователь выбирает режим работы программы. Если выбран режим создания эталона, за который отвечает модуль создания база данных (БД) эталонов, то программа обрабатывает и сохраняет входной сигнал с микрофона и выводит спектр на экран. Если же выбран режим распознавания, то программа обрабатывает результаты и сравнивает с заранее записанным эталоном в БД, сохраняет входной сигнал и переходит к его распознаванию с помощью вычисления первой и второй конечной разности полной фазовой функции, т.е. определяем количество звуков в данном слове, что видно из проделанного ранее моделирования, Определяем начало и конец слова с помощью выделения огибающей. Результат распознавания выводится на дисплей. На рис.25 представлен схематический вид программы. Схема изображена ниже.

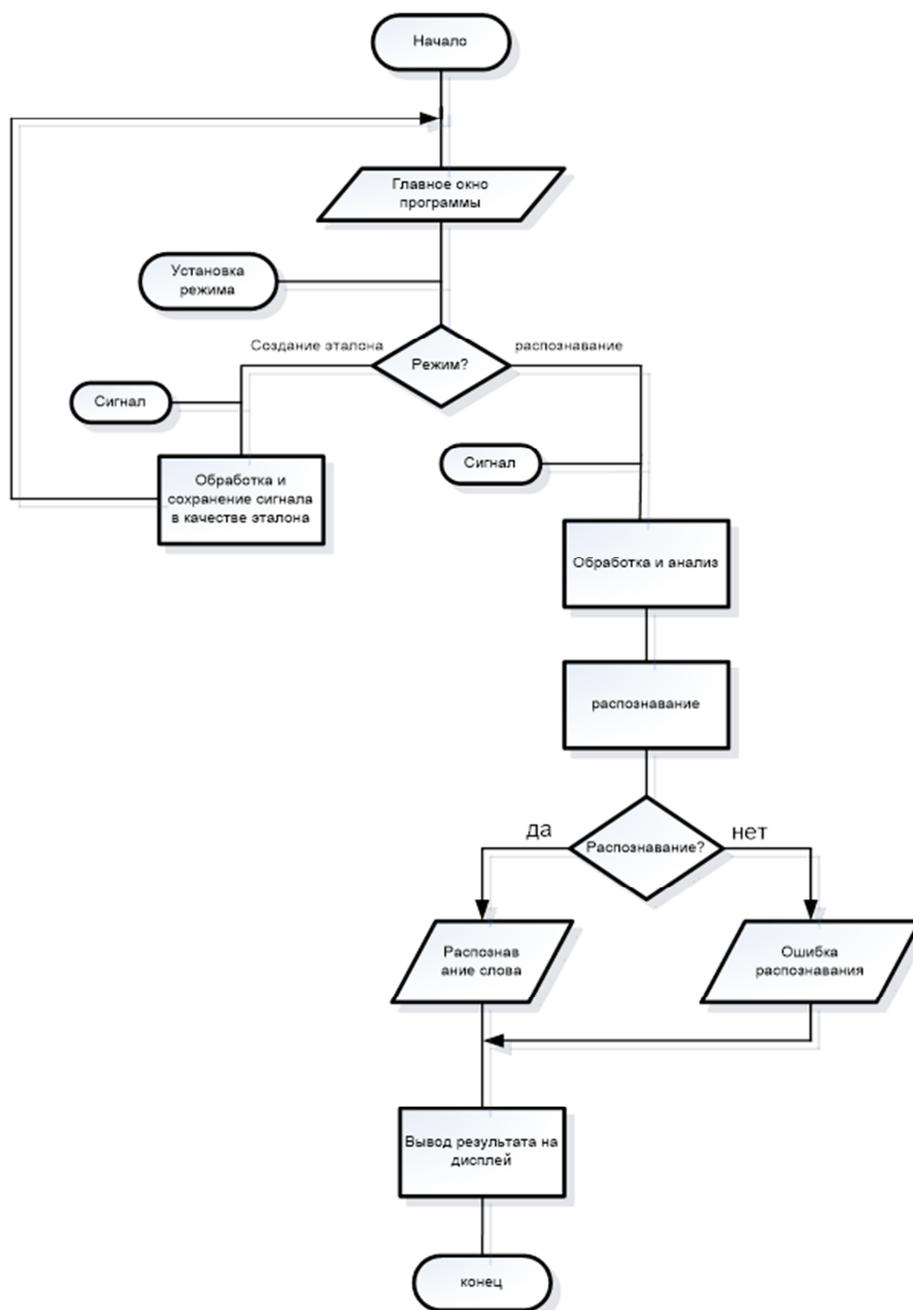


Схема 1. Алгоритм программы

Таблица 1

Данные по слову «Hello»

	Попытка 1	Попытка 2	Попытка 3	Попытка 4	Попытка 5
Человек 1	+	+	-	-	+
Человек 2	-	+	+	+	+
Человек 3	+	-	-	+	+
Человек 4	+	+	+	-	+
Человек 5	-	-	+	+	+

Итого получается, что процент распознавания слова «Hello» равен порядка 79%.

Данные по слову «Start»

	Попытка 1	Попытка 2	Попытка 3	Попытка 4	Попытка 5
Человек 1	-	-	+	+	+
Человек 2	+	+	+	+	+
Человек 3	+	+	-	-	+
Человек 4	+	+	+	-	-
Человек 5	-	+	-	+	+

Итого получается, что процент распознавания слова «Hello» равен порядка 83%.

Заключение

DTW алгоритмы являются очень полезными для распознавания отдельных слов в ограниченном словаре. Для распознавания беглой речи используются скрытые модели Маркова. Использование динамического программирования обеспечивает полиминальную сложность алгоритма: $O(n^2v)$, где n – длина последовательности, а v количество слов в словаре. DWT имеют несколько слабых сторон. Во-первых, $O(n^2v)$ сложность не удовлетворяет большим словарям, которые увеличивают успешность процесса распознавания. Во-вторых, трудно вычислить два элемента в двух разных последовательностях, если принять во внимание, что существует множество каналов с различными характеристиками. Тем не менее, DTW остается простым в реализации алгоритмом, открытым для улучшений и подходящим для приложений, которым требуется простое распознавание слов: телефоны, автомобильные компьютеры, системы безопасности и т.д.

Нейронные сети являются одним из наиболее перспективных методов распознавания речи. Данный метод позволяет подобрать топологию нейронной сети под решение конкретной задачи и позволяет оперировать с большим количеством слов в словаре без повышения трудоемкости процесса распознавания. Нейронные сети имеют гибкий аппарат обучения, позволяющий настроить сеть наилучшим образом для решения требуемой задачи.

Список литературы

- [1]. Фланаган Дж.Л. Анализ, синтез и восприятие речи / пер. с англ. А. А. Пирогова. М.: Связь, 1968. 397 с.
- [2]. Кузнецов В., Отт А. Автоматический синтез речи. Таллинн: Валгус, 1989. 135 с.

- [3]. Вишнякова О. А., Лавров Д. Н. Применение преобразования Гильберта-хуанга к задаче сегментации речи // Математические структуры и моделирование. 2011. вып. 24. С. 12–18
- [4]. Динамическое программирование. Режим доступа: <https://habrahabr.ru/post/113108/> (дата обращения 04.04.2016).
- [5]. Википедия. Нейронная система. Режим доступа: <https://ru.wikipedia.org/wiki/Нейросети> (дата обращения 04.04.2016).
- [6]. Михайлов В. Г., Златоустова Л. В. Измерение параметров речи. М.: Радио и связь, 1987. 168 с.