ФИЗИКА-МАТЕМАТИКА ҒЫЛЫМДАРЫ

УДК 01.04

Амиргалиев Е.Н.
Университет имени Сулеймана Демиреля
Мусабаев Р.Р., Мусабаев Т.Р.
Институт информационных и вычислительных технологий КН МОН РК

О ПОДХОДЕ ПАРАМЕТРИЗАЦИИ И СЕГМЕНТАЦИИ РЕЧЕВОГО СИГНАЛА

Аннотация. В статье описан способ параметризации речевого сигнала с помощью алгоритмов анализа, синхронизированных с частотой основного тона. Результаты параметризации использованы для сегментирования речевого сигнала и выделения сегментов со стабильными спектральными характеристиками. Описаны этапы достижения сегментации слитной речи, рассмотрены различные виды устройств распознавания речи, а также области их применения, перспективы их использования в организации процесса речевой коммуникации человека с системами искусственного интеллекта в связи с повышением на них коммерческого спроса. Также авторами приводится анализ основных проблем в области распознавания слитной речи. Это распознавание и смысловая интерпретация слитной речи для устного диалога человека и ЭВМ на естественном языке. В статье представлены различные методы обработки и записи, а также полученные в ходе исследований результаты в виде графических изображений речевого сигнала и сигнала основного тона.

Принципиальное предположение, которое делается в современных программах речевого распознавания, что речевой сигнал рассматривается как стационарный, а размеры сегмента берутся случайно и одной величины. Типичная величина одного сегмента — 25 мс.

Авторами статьи предложено формировать речевые сегменты посредством выделения в сигнале основного тона положительных экстремальных значений и проведения через них границы сегментов, а затем синхронизировать выделенные сегменты по времени с речевым сигналом.

Ключевые слова: сегментирование речевого сигнала, частота основного тона, параметризация речи, фильтрация сигнала, методы, запись, ЭВМ, речевая коммуникация.

ВВЕДЕНИЕ

На современном этапе к работам в сфере распознавания речи человека объектами искусственного интеллекта принадлежат как труды, связанные с непосредственным применением распознавателей обособленных слов в деятельности коммерческих предприятий и государственном секторе, так и прикладные научные исследования, ориентированные, прежде всего на создание универсальных распознавателей сложных предложений. В настоящее время имеет место всплеск коммерческого спроса на устройства распознавания речи, а также увеличивается область их применения, что свидетельствует о широких перспективах в организации процесса речевой коммуникации человека с системами искусственного интеллекта. Распознавание речи выступает неотъемлемым элементом научно-исследовательских работ, проводимых в более широкой области знаний, получившей название обработки речи. Помимо вопросов, связанных с проблематикой

распознавания речи, данная сфера знаний включает в себя идентификацию говорящих при помощи систем искусственного интеллекта, машинный синтез речи и воспроизведение хранящихся в ЭВМ речевых ответов, машинный анализ степени физического и психологического состояния говорящего, эффективную трансляцию публичных выступлений и диалогов, а также диагностику речевых дефектов и оказание помощи больным, страдающим нарушениями речи [1]. В настоящей статье предметом для рассмотрения являются только те вопросы, которые затрагивают проблемы нахождения оптимальных результатов параметризации и сегментирования речевого сигнала для формирования цифрового голосового образа личности или применения в системе автоматического распознавания речи.

На сегодняшний день в области распознавания слитной речи существует несколько проблем. Одна из таких, научить ЭВМ распознавать в слитной речи гласные и согласные фонемы. Для решения этой проблемы авторами было предложено использовать кратковременные алгоритмы анализа, синхронизированные с частотой основного тона. В статье будет рассмотрен именно данный подход, для создания оптимальной параметризации и сегментирования речевого сигнала. В дальнейшем, синхронизированная с частотой основного тона параметризация сигнала может помочь в формировании и анализе цифрового голосового образа личности.

Сформулированная авторами задача сегментации речевого сигнала относится к основным проблемам анализа, распознавания и интерпретации речевых сигналов, т.к. подчинена проблеме распознавания слитной речи, имеющей большую практическую значимость. Это – распознавание и смысловая интерпретация слитной речи для устного диалога человека и ЭВМ на естественном языке. [2]В практическом плане решение этой задачи может явиться ключом к достижению полной эффективности речевого ввода информации в ЭВМ.

В настоящее время общие математические модели речевого сигнала еще не найдены. Формулируются лишь частные модели, ориентированные на постановку и решение частных задач распознавания речи. Примерами таких задач являются распознавание отдельно произносимых слов для одного диктора, распознавание диктора по парольной фразе, определение функционального состояния одного данного человека по его голосу. [2].

Получение речевого сигнала и его декомпозиция. В первую очередь авторами был записан эталон речи женского типа голоса, дискретизированный с частотою 44 КГц и с разрядностью 16 бит. Запись эталона осуществлялась посредством применения микрофона и ларингофона позволяющего получать звук непосредственно от вибраций голосовых связок.

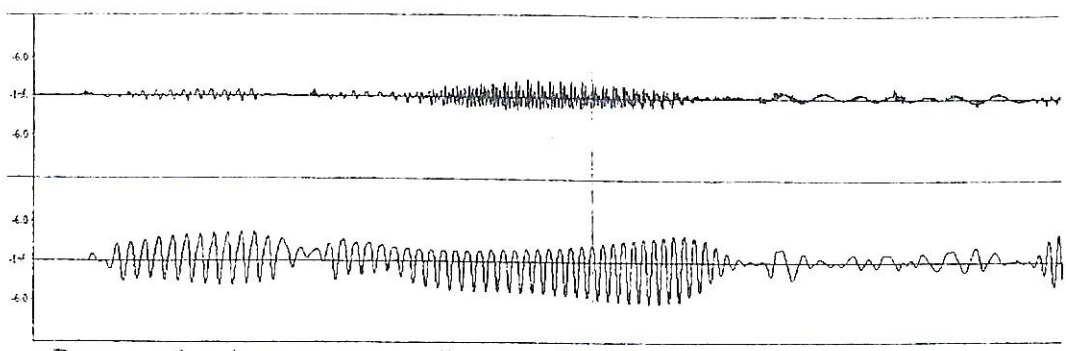

Рисунок 1 – Агрегированный речевой сигнал и сигнал основного тона

$$g_i = \frac{1}{K+1} \sum_{j=-K}^{0} f_{i+j} (i = 1+K, 2+K, \ldots, N) \tag{3}$$

Когда мы производим сглаживание, нужно определить *вес* каждой точки в соответствии с ее значимостью. Это можно записать следующим выражением:

$$g_i = \sum_{j=-K}^{K} w_i f_{i+j} (i = 1+K, 2+K, 3+K, \ldots, N-K) \tag{4}$$

Чтобы не исказить величину усредняемой функции, примем следующее условие:

$$\sum_{j=K}^{K} w_i = 1 \tag{5}$$

где wi – функция, дающая вес точкам. В качестве весовой функции обычно используется функция распределения Гаусса.[6] На рисунке 2 продемонстрированы результаты обработки исходного сигнала основного тона КИХ-фильтром.

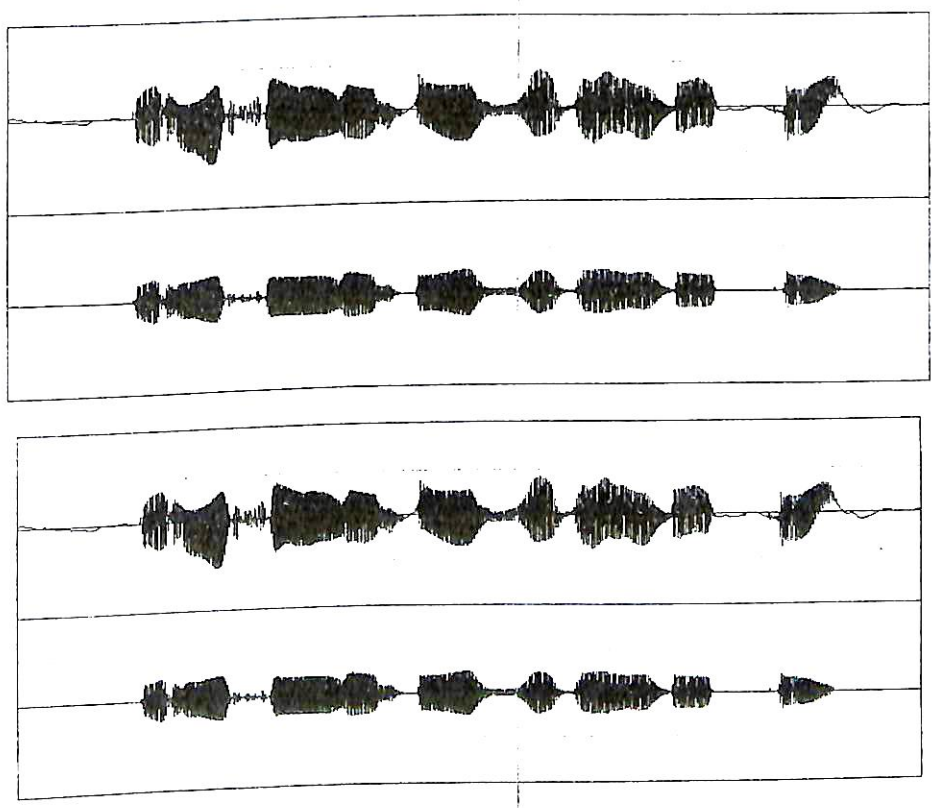

Рисунок 2 – Исходный сигнал основного тона (вверху), тот же сигнал обработанный КИХ-фильтром (внизу).

Для начала необходимо синхронизировать по времени речевой сигнал с частотой основного тона. Акустическим коррелятом тона является так называемая частота основного тона (ЧОТ), определяемая как частота вибрации голосовых связок. С акустической точки зрения ЧОТ — это первая гармоника речевого сигнала. [3] Частота колебаний голосовых связок, от которой зависит высота тона (чем чаще колеблются связки, тем выше частота основного тона голоса), определяется напряженностью, длиной и общей массой голосовых связок. [4] Для синхронизации речевого сигнала с сигналом основного тона можно использовать технику одновременной записи сигналов с микрофона и ларингофона. В нашем случае частота основного тона задаёт период повторения звуковых колебаний, что может существенно помочь при классификации речевого сигнала.

В следующем шаге производится декомпозиция записанных сигналов с целью последующей обработки и применения различных фильтров к каждому из сигналов. Фактически, в любом речевом сигнале, записанном в рабочем помещении, может содержаться много мешающей информации: различный шум, присутствие высоких и низких частот, имеющих высокую частоту колебания и т.д. Стоит обратить внимание на то, что у каждого говорящего базовая частота основного тона индивидуальна и обусловлена особенностями строения гортани. В среднем для мужского голоса она составляет от 80 до 210 Гц, для женского — от 150 до 320 Гц.[5] Поэтому частоты выше или ниже этого диапазона считаются излишними для решения задачи выделения ЧОТ. Присутствие этой избыточной информации может исказить классифицирование исходного речевого сигнала. Избавиться от этой информации можно применением фильтра с конечной импульсной характеристикой (не рекурсивный фильтр, КИХ-фильтр).

Сглаживание речевых сигналов. Если мы хотим увидеть приблизительную динамику изменения сигнала, необходимо сделать его «гладким», удалив незначительный шум, входящий в сигнал, и устранив мелкие колебания сигнала. Этот вид обработки называется сглаживанием сигнала. В этот вид обработки может входить операция, которая называется скользящим усреднением, являющаяся элементарной формой КИХ-фильтра. То есть, берем некоторую область до и после рассматриваемой точки и, учитывая численные значения измерений, входящих в эту область, вычисляем среднее значение.

Это определяется соотношением:

$$g_i = \frac{1}{K+1}(f_{i-K} + f_{i-K+1} + \ldots + f_i) \quad (1)$$

где N точек измерений цифрового сигнала $(f_1, f_2, \ldots, f_N)$; $f_i, i = 1, 2, 3, \ldots, N$.

Для нахождения скользящего среднего в окрестности рассматриваемой точки i берем среднее арифметическое от K предыдущих и последующих точек, включая точку i. Кстати, обратим внимание на то, что на первых и на последних точках i оси абсцисс невозможно вычислить значение сглаживания. Область, где это возможно сделать, определяется следующим образом:

$$i = 1+K, 2+K, \ldots, N-K. \quad (2)$$

С использованием знака суммы соотношение (1) записывается в виде:

Надо упомянуть, что применение КИХ-фильтра для сглаживания полученных сигналов было выбрано не случайно. На самом деле существует два основных типа цифровых фильтров: фильтры с конечной импульсной характеристикой (КИХ) и фильтры с бесконечной импульсной характеристикой (БИХ). Если речь идет об обработки речевого сигнала, то эта классификация относится к импульсным характеристикам фильтров. Изменяя веса коэффициентов и число звеньев КИХ-фильтра, можно реализовать практически любую частотную характеристику.[7] КИХ-фильры обладают рядом полезных свойств:

- КИХ-фильтры устойчивы;
- КИХ-фильтры при реализации не требуют наличия обратной связи;
- Фаза КИХ-фильтров может быть сделана линейной.[8]

После того как исходный звуковой сигнал и частота основного тона сглажены, необходимо предварительно обработать речевой сигнал для получения множества спектральных векторов, характеризующих этот сигнал. Для чего это делается? Поскольку речь является нестационарным процессом (т.е. его спектральные характеристики относительно непостоянные), то ее принято анализировать на коротких участках (10 — 30 мс), где спектрально-корреляционные характеристики остаются примерно постоянными. Принципиальное предположение, которое делается в современных программах речевого распознавания, что речевой сигнал рассматривается как стационарный, а размеры сегмента берутся случайно и одной величины. Типичная величина одного сегмента — 25 мс.

Автором статьи, было предложено формировать эти сегменты посредством выделения в сигнале основного тона положительных экстремальных значений и проведения через них границы сегментов. А затем синхронизировать выделенные сегменты по времени с речевым сигналом. Данный прием сможет обеспечить увеличение точности при параметризации речевого сигнала и выбрать оптимальное значение размера сегмента. Нахождение экстремальных значений является простой математической задачей и в данной статье рассматриваться не будет. Сообщим только, что поиск экстремальных значений осуществлялось простым перебором с помощью сравнительного метода ближайшего соседа по значениям оси ординат.

При помощи описанного выше метода были получены сформированные оптимальные сегменты речевого сигнала со средним значением сегмента по времени в 9 мс, которая зависит от частоты дискретизации сигнала основного тона. Результат можно увидеть на рисунке 3.

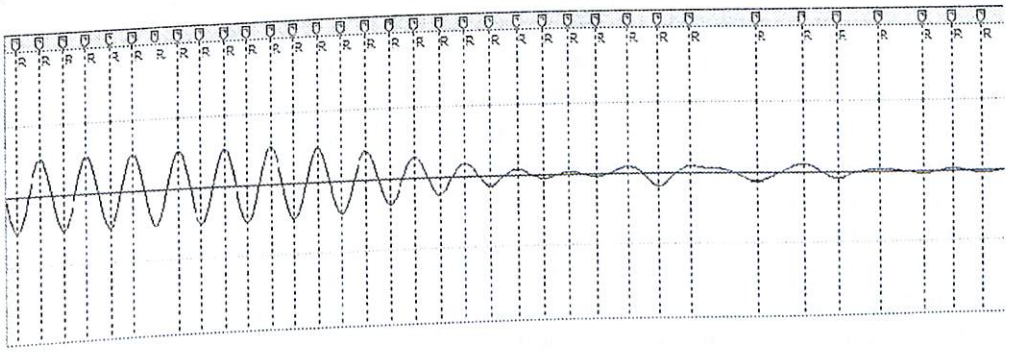

Рисунок 3 – Выделенные полусегменты сигнала основного тона по его частоте дискретизации.

Параметризация речевого сигнала. В основе большинства методов обработки речи лежит предположение о том, что свойства речевого сигнала с течением времени медленно изменяются. Это предположение приводит к методам кратковременного анализа, в которых сегменты речевого сигнала выделяются и обрабатываются так, как если бы они были короткими участками отдельных звуков с отличающимися свойствами. Процедура повторяется так часто, как это требуется. Результатом обработки на каждом сегменте является число или совокупность чисел.[9] Для классификации речевого сигнала на гласные и согласные могут быть использованы методы обработки во временной области. Под областью здесь подразумевается ранее выделенные сегменты речевого сигнала, синхронизированные с частотой дискретизации основного тона. Для обработки временного сегмента речевого сигнала были использованы кратковременные функции среднего числа переходов через ноль и энергии сигнала. Эти методы являются наиболее важными параметрами для голосовой классификации и очень часто используются в автоматических системах распознавания речи.

Функция среднего числа переходов через ноль. При обработке сигналов в дискретном времени считают, что если два последовательных отсчета имеют различные знаки, то произошел переход через нуль. Частота появления нулей в сигнале может служить простейшей характеристикой его спектральных свойств. Рассмотрим способ вычисления этой величины. Определим среднее число переходов через нуль:

$$Z_n = \sum_{m=-\infty}^{\infty} \left| sgn[x(m)] - sgn[x(m-1)] \right| w(n-m), \tag{6}$$

где

$$sgn[x(n)] = \begin{cases} 1, x(n) \geq 0, \\ -1, x(n) < 0 \end{cases} \tag{7}$$

и

$$w(n) = \begin{cases} \dfrac{1}{2N}, 0 \leq n \leq N-1, \\ 0, в противном случае. \end{cases} \tag{8}$$

Модель речевого образования предполагает, что энергия вокализованных сегментов речевого сигнала концентрируется на частотах ниже 3 кГц, что обусловлено убывающим спектром сигнала возбуждения, тогда, как для невокализованных сегментов большая часть энергии лежит в области высоких частот. Поскольку высокие частоты приводят к большему числу переходов через нуль, а низкие – к малому, то существует жесткая связь между числом нулевых пересечений и распределением энергии по частотам.[9]Надо заметить, что число нулевых пересечений для временных сегментов характеризующих гласные и согласные звуки, также различен. Согласные – это звуки, при произнесении которых воздух встречает на своем пути преграду. Щель и смычка – это два основных способа образования согласных. Вид преграды определяет характер согласного звука. Наличие шума – также один из отличительных признаков согласных. Поэтому согласные по своей природе относятся к высоким частотам, имеющим большее число переходов через нуль. В то же время гласные звуки, при произношении которых воздух свободно проходит через ротовую

полость, не встречая на своем пути преграды, можно отнести к низким частотам с меньшим числом переходов через ноль.[10] На рисунке 4, можем наблюдать, что гласные имеют «сглаженную» и относительно низкую амплитуду на графике, реализующего функцию среднего числа перехода через ноль. А согласные в противоположность гласным имеют «неровную» и относительно высокую амплитуду колебания.

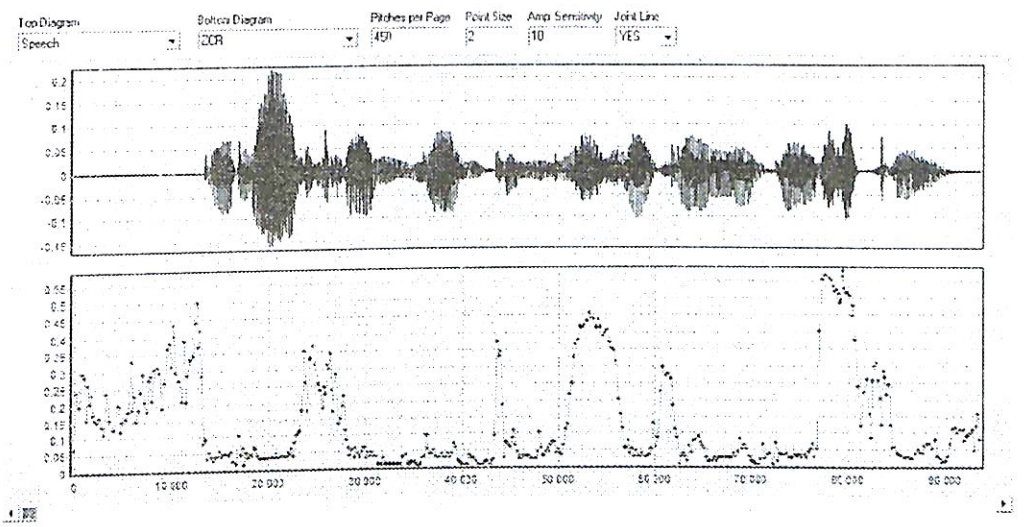

Рисунок 4 – Параметризованный речевой сигнал с помощью функции среднего числа
переходов через ноль

Кратковременная энергия сигнала. Как отмечалось выше, амплитуда речевого сигнала существенно изменяется во времени. В частности, амплитуда невокализованных сегментов речевого сигнала значительно меньше амплитуды вокализованных сегментов. Подобные изменения амплитуды хорошо описываются с помощью функции кратковременной энергии сигнала. В общем случае определить функцию энергии можно как

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2, \quad (9)$$

Это выражение может быть переписано в виде

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2, \quad (10)$$

где

$$h(n) = w^2(n). \quad (11)$$

Сигнал x2(n) в этом случае фильтруется с помощью линейной системы с импульсной характеристикой h(n).

Для описания быстрых изменений амплитуды желательно иметь узкое окно (короткую импульсную характеристику), однако слишком малая ширина окна может

привести к недостаточному усреднению и, следовательно, к недостаточному сглаживанию функции энергии.[9] Поэтому размер речевого сегмента для обработки этой функцией был выбран не случайно, а синхронизирован с частотой основного тона. На рисунке 5 изображен результат использования функции кратковременной энергии сигнала. В противоположность функции среднего числа переходов через ноль, гласные на графике функции энергии сигнала будут иметь высокую амплитуду.

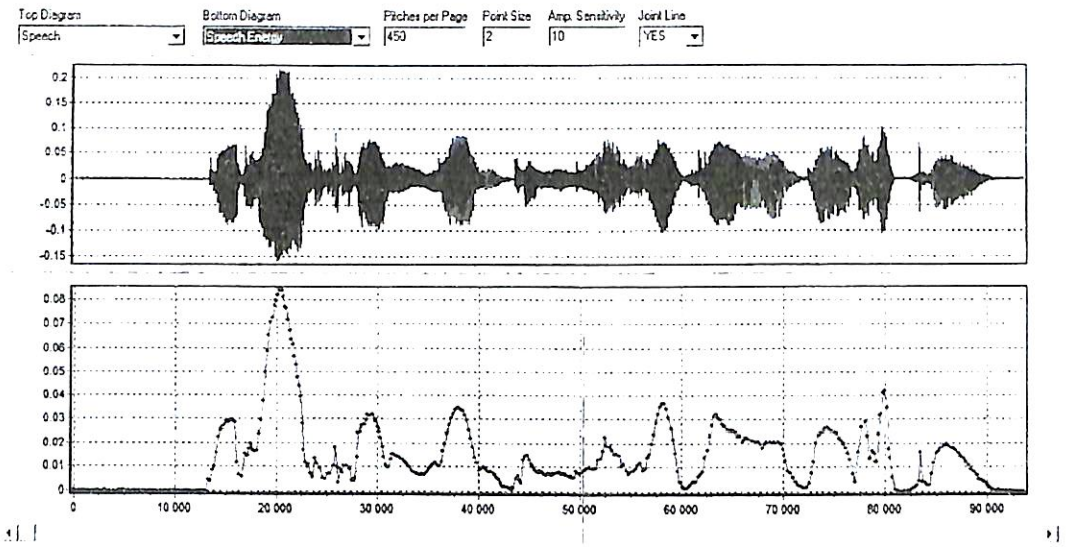

Рисунок 5 – Параметризованный речевой сигнал с помощью функции энергии.

Речевой сигнал состоит из квазистационарных участков, соответствующих голосовым и шипящим фонемам, перемежаемых участками со сравнительно быстрыми изменениями спектральных характеристик сигнала (межфонемные переходы, взрывные и смычные фонемы, внутрисловные переходы речь-пауза). В пределах стационарных участков значительную роль для анализа речи играют спектральные особенности сигнала, определяемые передаточной характеристикой речевого тракта, изменяющейся в процессе артикуляции. Можно сказать, что речевой сигнал характеризуется нелинейными флуктуациями различных масштабов. Если в качестве структурных единиц речи рассматривать фонемы, то задача сегментации сводится к обнаружению межфонемных переходов.[11]

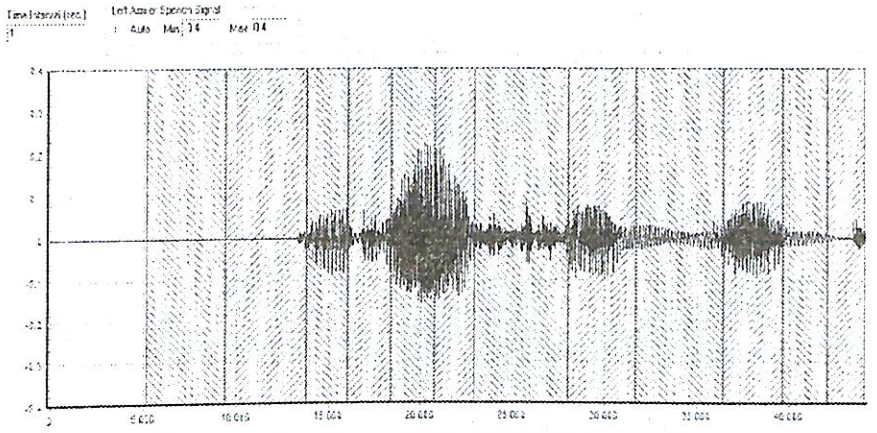

Рисунок 6 – Сегментированный речевой сигнал

Для определения границ квазистационарных участков можно использовать результаты ранее проведенной параметризации. Как видно на графиках кратковременных функций среднего числа переходов через ноль и энергии сигнала, можно заметить разницу в высоте амплитуды и ее «сглаженности». Для определения границ используется метод перебора амплитудных значений с последующим нахождением наибольшего разностного значения по оси ординат между двумя ближайшими значениями амплитуд. Если такое разностное значение было найдено, то можно предположить, что происходит быстрое изменение спектральных характеристик речевого сигнала. Можно также отметить, что сглаживание результатов параметризации фильтром с конечной импульсной характеристикой перед началом процедуры сегментирования речевого сигнала может удалить ненужные мелкие колебания сигнала и повысить точность сегментирования. На рисунке 6 представлен график сегментированного речевого сигнала на основе результатов параметризации функцией кратковременной энергии сигнала.

В дальнейшем, на основе результатов сегментирования можно классифицировать исходный речевой сигнал на гласные и согласные фонемы с помощью применения кластерного анализа.

Список использованной литературы

1 Ли. У – Методы автоматического распознавания речи. Том 1. 1983. – С. 7-18

2 Винцюк Т.К., Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наук, думка, 1987. – 13–22 с.

3 Кодзасов, С. В., Кривнова, О. Ф. Общая фонетика. М., РГГУ: 2001, стр. 106.

4 http://fonetica.philol.msu.ru/nn/n15.htm

5 Ashby, Michael, & John Maidment. Introducing Phonetic Science. Cambridge: CUP, 2005

6 Юкио Сато, Обработка сигналов. Первое знакомство. – Додэка, 2002. – 29–32 с.

7 Уолт Кестер, Цифровая обработка сигналов, Глава 6. – 6 с.

8 Шишков А.Н., Цифровые сигнальные процессоры. – 24 с.

9 Рабинер Л.Р., Шафер.Р.В., Цифровая обработка сигналов. – М.: Радио и связь, 1981. – 112–121с.

10 http://russkiy-na-5.ru/articles/157

11 Вишнякова О.А., Лавров Д.Н., Автоматическая сегментация речевого сигнала на базе дискретного вейвлет-преобразования. – Математические структуры и моделирование, 2011, вып. 23. – 43–48 с.

Amirgaliev Y.N.
professor, Suleyman Demirel University, Kazahstan, Kaskelen
Musabayev R.R., Musabayev T.R.
MES RK SK Institute of Information and computing technologies, Kazakhstan, Almaty

ABOUT AN APPROACH PARAMETRIZATION AND SEGMENTATION OF THE SPEECH SIGNAL

Abstract: The article describes how the parameterization of the speech signal by analysis algorithms that are synchronized with the pitch frequency. The results are used for parameterization of the speech signal segmentation and allocation segments with stable spectral characteristics. Stages achieve continuous speech segmentation are considered different types of speech recognition devices, and their applications, prospects for their use in the organization of human speech communication process with the artificial intelligence systems in connection with increasing commercial demand for them. The author also analyzes the main problems in the field of continuous speech recognition. This recognition and semantic interpretation of continuous speech for the oral dialogue and human computer in natural language. The article presents various methods of processing and recording, as well as obtained in the course of research results in the form of graphic images of the speech signal and the pitch signal.

The fundamental assumption is that in today's speech recognition programs, that the speech signal is considered as stationary, and segment sizes are taken randomly and the same values. A typical value of one segment - 25 msec.

The authors proposed to form segments by separating the speech signal into a pitch and positive extreme values of the boundary segments through them, and then synchronize the time segments allocated to the speech signal..

Key words: segmenting the speech signal, pitch frequency, parameterization speech signal filtering techniques, recording, computers, voice communication.

ӘмірғалиевЕ.Н.
т.ғ.д., профессор,Сулейман Демирельат.университет, Қазақстан, Қаскелен
Мусабаев Р.Р., Мусабаева Т.Р.
ҚР БҒМ ҒК Ақпараттық және есептеуіш технология институты, Қазақстан, Алматы

СӨЙЛЕУ СИГНАЛЫНЫҢ ПАРАМЕТРИЗАЦИЯЛАУ МЕН СИГМЕНТТЕУІНІҢ БІР ЖАҚТЫЛЫҒЫ

Аңдатпа: Мақалада негізгі ырғақтың жиілігімен синхрондалған талдау алгоритмінің көмегімен сөйлеу сигналдарын параметрлеу әдісі көрсетілген. Параметрлеу нәтижесі сөйлеу сигналдарын сегменттеуге және тұрақты спектрлік сипатамалы сегменнттерді бөлуге қолданылған. Жалғасқан сөйлеулерді сегменттеу кезеңдері сипатталған.

Кілт сөздер: сөйлеу сигналын сегменттеу, негізгі ырғақтық жиілік, сөйлеуді параметрлеу, сигналды фильтрлеу.