Алгоритм автоматической сегментации фрагментов речи в системе «человек – машина»

Algorithm for automatic segmentation of speech fragments in the "man – machine" system

Смирнов / Smirnov V.
Виктор Михайлович
(Vm_smir@mail.ru)
кандидат технических наук, доцент.
ФГАОУ ВО «Санкт-Петербургский государственный университет аэрокосмического приборостроения» (ГУАП), старший научный сотрудник.
г. Санкт-Петербург

Филатов / Filatov V.
Владимир Николаевич
(filbait@mail.ru)
кандидат технических наук.
ГУАП, доцент.
г. Санкт-Петербург

Ключевые слова: речь – speech; фонема – phoneme; сегментация – segmentation; вейвлет-преобразование – wavelet-transformation; кратномасштабный анализ – discrete analise;вейвлет Добеши – wavelet Daubechies.

В статье рассматриваются возможности и перспективы использования устной речи в современных системах «человек – машина» на обитаемых космических станциях. Для сегментации речевого сигнала с целью выделения фрагментов фонем русского языка предлагается использование кратномасштабного вейвлет-преобразования Добеши. Для более точного определения границ предлагается проводить постобработку для обнаружения «сильных» и «слабых» границ. Приводятся результаты экспериментальных исследований предложенного алгоритма сегментации различных по сложности произношения слов.

The possibilities and prospects of spoken language usage in modern "man-to-machine" systems intended for the inhabited space stations are discussed in the article. For the segmentation of speech signal aimed to detect fragments of phonemes of the Russian language, the multiple-scale Daubechies wavelet transform is proposed. For more accurate definition of boundaries it is proposed to carry out post-processing targeted to identify "strong" and "weak" boundaries. The results of experimental studies of the proposed algorithm for segmentation of words with different complexity of pronunciation are presented.

Введение

Наша речь создается с помощью сложнейшего инструмента – голосового тракта, возможности которого человечество пытается скопировать и повторить несколько десятилетий. И дело не только в том, что человеческий голос остается непревзойденным по своим качествам (тембру, возможности передачи тончайших нюансов звучания), но и в том, что он является средством передачи вербальной (словесной) информации. Для быстрой передачи смысловой информации в процессе эволюции был сформирован особым образом закодированный и структурированный акустический (речевой) сигнал.

Устная речь – это наиболее продуктивный, естественный и удобный способ передачи информации, поэтому с момента появления первых электронно-вычислительных машин одним из наиболее важных вопросов был процесс взаимодействия человека с машиной. В современных компьютерных системах все больше внимания уделяется вопросам построения интерфейса речевого ввода-вывода, поскольку его потенциальная эффективность основана на практически неограниченных возможностях формулировки на естественном языке всевозможных задач в самых разных областях человеческой деятельности.

Потребность в подобного рода системах появилась и в космической отрасли, а точнее в обитаемых космических аппаратах. Первый опыт голосового общения робота Kirobo с японским космонавтом Коити Вакаты был успешно проведен на международной космической станции в 2013 году. В августе 2019 года российский человекообразный робот «Федор» также побывал на МКС, выполняя функции голосового помощника космонавтов. Эти эксперименты показали целесообразность выполнения коммуникативных функций роботизированными системами в космосе.

Таким образом, создание и развитие систем голосового управления различными устройствами космической станции, голосового доступа к справочным системам, голосового общения с роботом-помощником, позволяющих упростить работу космонавтов, увеличить ее эффективность и безопасность, можно считать актуальной и насущной задачей.

Методы распознавания речи

Одной из важнейших задач в системах автоматической обработки речи является задача сегментации в соответствии с фонетической транскрипцией языка,

которая, в свою очередь, предполагает решение нескольких задач. Прежде всего, поскольку звук – это акустический сигнал, то есть механическое колебание упругой среды, необходимо преобразовать колебания воздуха в электрические сигналы при помощи микрофона, отфильтровав при этом помехи и шумы. Далее уже электрический сигнал необходимо представить в цифровой форме, ввести его в компьютер для дальнейшей обработки. На сегодняшний день эти задачи не представляют трудности с точки зрения технической реализации. В компьютер можно ввести информацию как об амплитуде сигнала, так и о его спектральном составе.

Следующая задача – выделение из речевого потока лингвистических конструкций – осложняется тем, что человек произносит слова, не отделяя слоги друг от друга, и с минимальными паузами между словами. В реальной жизни речь представляет собой сплошной поток звуков. В процессе формирования непрерывной речи звуки соответствующие одним и тем же фонемам, могут изменяться при соединении с другими звуками.

Кроме того, звуки непрерывной речи имеют постоянно изменяющийся спектр гармонических частот, а также шум. Громкость и темп речи также постоянно изменяются, одна и та же фраза, сказанная разными людьми, или даже одним человеком в зависимости от его психологического состояния, может иметь разную спектрально-временную окраску.

Несмотря на разницу в произношении в различных языках существует некоторый типизированный простейший звук речи, называемый фонемой. Фонема – лингвистическая единица речи. При этом все звуки речи можно разбить на две большие группы: гласные и согласные, которые отличаются принципами формирования. В русском языке используются 42 фонемы, из них 6 гласных и 36 согласных [1]. Согласные звуки несут в основном смысловую нагрузку в тексте речи, а гласные – эмоциональную. Таким образом непрерывный поток речи сводится к последовательности произношения во времени отдельных фонем и их акустического окружения.

Для того чтобы выделить из оцифрованного звука лингвистические единицы речи, применяются различные математические методы в сочетании со специальным программным обеспечением. Как правило, используются два принципиально разных подхода:

- распознавание голосовых меток;
- распознавание лингвистических единиц речи.

Первый подход предполагает распознавание фрагментов речи по заранее записанному образцу. Этот подход широко используется в системах, предназначенных для исполнения заранее записанных речевых команд.

Второй подход предполагает выделение из потока речи отдельных лингвистических единиц речи – фонем и аллофонов, которые затем объединяются в слоги и морфемы, то есть производится сегментация речевого сигнала.

Необходимо сказать, что сегментация, сама по себе, является своего рода задачей распознавания. Особенность сегментации состоит в том, что она может решаться для каждого речевого сигнала отдельно, то есть необходимо распознать границы конкретных объектов – аллофонов данного речевого сигнала, а не сами аллофоны. Это дает возможность идентифицировать границы аллофонов в словах, произнесенных с различной интонацией, так как для сегментации используется только информация, содержащаяся в данном сигнале.

Здесь, как и в общей проблеме распознавания, возможно несколько вариантов. Один из вариантов – использование ручной сегментации. Однако ручная сегментация требует значительных затрат сил и времени, поскольку в речевом сигнале практически нет пауз между словами. Кроме того, процесс коартикуляции, возникающий на границе последовательно производимых звуков и существенно облегчающий правильное восприятие, понимание речи, затрудняет задачу сегментации.

Ко второму варианту относятся способы автоматической сегментации. Среди наиболее известных можно выделить следующие:

– способы, основанные на применении значений кратковременной энергии речевого сигнала и числа переходов через ноль [2];

– способы, основанные на применении значений информационной энтропии [3, 4];

– способы, основанные на применении мелчастотных кепстральных коэффициентов [5,6];

– способы, основанные на статических моделях [7,8].

Еще одним из способов сегментации, учитывающим, то, что речевой сигнал характеризуется нелинейными флуктуациями различных масштабов, является кратномасштабный анализ и вейвлет-преобразование [9].

Выбор вейвлет-преобразования обусловлен несколькими причинами.

С точки зрения речевого сигнала фонемам соответствует квазистационарный участок, а акустическому окружению – участки со сравнительно быстрыми изменениями спектральных характеристик сигнала – межфонемные переходы, взрывные и шипящие фонемы, внутрисловные переходы речи (паузы) [10]. В пределах квазистационарных участков значительную роль для анализа речевого сигнала играют спектральные особенности сигнала (распределение формантных частот), определяемые передаточной характеристикой речевого тракта, изменяющейся в процессе артикуляции. Можно сказать, что речевой сигнал характеризуется нелинейными флуктуациями различных масштабов. Поэтому весьма эффективным для анализа речевого сигнала представляется применение кратно-масштабного анализа и вейвлет-преобразования.

Применение вейвлет-преобразования для сегментации речевого сигнала

Вейвлет-преобразование хорошо согласуется с моделью слуховой системы человека. Восприятие частотных различий в звуковом сигнале согласно

«теории места» можно представить, как анализ сигнала на базилярной мембране многоканальным банком фильтров, имеющих полосу пропускания, равную ширине критических полос слуха. Критическая полоса слуха до частоты 1 кГц считается постоянной (порядка 170 Гц), а дальше растет и составляет 17% от центральной частоты фильтра. Слух фиксирует изменения спектральных составляющих в логарифмическом масштабе. Он распознает звуки по изменению их спектров во времени, т.е. по фонетическому содержанию речи и по приращению скорости изменения спектральных составляющих на заданной частоте.

Любая дискретная функция, в том числе и речевой сигнал, может быть представлена в виде линейной комбинации вейвлет-функций на различных масштабах (уровнях декомпозиции) и скейлинг-функции на самом большом масштабе разрешения:

$$S(t)=\sum_{k=0}^{\frac{N}{2^n}-1} a_{j_n,k}\,\varphi_{j_n,k}(t)+\sum_{j=1}^{n}\sum_{k=0}^{\frac{N}{2^j}-1} d_{j,k}\,\psi_{j,k}(t)=$$

$$=\sum_{k=0}^{\frac{N}{2^n}-1} a_{j_n,k}\,\varphi_{j_n,k}(t)+\sum_{k=0}^{\frac{N}{2^j}-1} d_{j,k}\,\psi_{j,k}(t)+\sum_{k=0}^{\frac{N}{2^{(j-1)}}-1} d_{(j-1),k}\,\psi_{(j-1),k}(t)+\ldots$$

$$\ldots+\sum_{k=0}^{\frac{N}{2^1}-1} d_{1,k}\,\psi_{1,k}(t), \qquad (1)$$

где N – число отсчетов фрейма; n – количество уровней декомпозиции; k – диапазон сдвигов; $a_{j,k}=\int S(t)\varphi_{j,k}(t)dt$ – коэффициенты аппроксимации на j-ом уровне декомпозиции, связанные со средними значениями речевого сигнала; $d_{j,k}=\int S(t)\psi_{j,k}(t)dt$ – детализирующие коэффициенты на j-ом уровне декомпозиции, связанные с флуктуациями речевого сигнала; $\varphi_{j,k}=2^{j/2}\varphi(2^j t-k)$ и $\psi_{j,k}=2^{j/2}\psi(2^j t-k)$ – масштабированные и смещенные версии скейлинг-функции (масштабной функции) φ, имеющей низкочастотный характер, и "материнского вейвлета" ψ.

Обычно в качестве параметра, определяющего выбор вида материнского вейвлета, выступает внешнее сходство вида исследуемого сигнала и функции преобразования. Исходя из этого, при исследовании в качестве материнской вейвлет-функции использованы вейвлеты Добеши, основными свойствами которых являются:

1) функции имеют конечное число нулевых значений, т.е. система вейвлетов Добеши обладает свойствами гладкости и исключения моментов;

2) функции обладают свойствами компактности носителя (т.е. быстро нарастают и быстро спадают) и ортогональности, что обусловливает возможность точного восстановления произвольного сигнала;

3) вейвлеты имеют как вейвлет-функцию, так и скейлинг-функцию, что делает возможным кратномасштабный и быстрый вейвлет-анализ.

При использовании вейвлет-коэффициентов в качестве признаков, описывающих речевой сигнал, необходимо определить число уровней декомпозиции, соответствующих размеру анализируемого частотного диапазона.

Частотный диапазон акустических колебаний, в котором мы воспринимаем их как звук, составляет от 20 Гц до 20 кГц. Однако специфика формирования речевого сигнала приводит к тому, что речевые звуки занимают примерно одну треть этого диапазона. Дело в том, что основная частота колебаний голосовых связок, которая определяет нижнюю границу диапазона, равна у мужских голосов в среднем 110 Гц, у женских – 220 Гц, у детских – 300 Гц, а максимальная частота речевых звуков близка к 6500 Гц. В работах по речеобразованию показано, что вполне удовлетворительная разборчивость речи может быть обеспечена в диапазоне частот от 500 до 4000 Гц, а с учетом обертонов верхняя граница может подниматься до 8 кГц.

Как известно, вейвлет-преобразование не дает частотно-временного представления сигнала. Вместо этого используется пространство “время-масштаб”. При этом значения масштабных коэффициентов можно перевести в квазичастоты и получить результат в частотном представлении [11]. Квазичастота j-го уровня декомпозиции определяется как

$$F_j=\frac{F_c M}{a}, \qquad (2)$$

где F_c – частота центрального всплеска вейвлета; M – число отсчетов в секунду; a – масштабный коэффициент.

Рассчитанные в MatLab частоты центрального всплеска для различных базовых вейвлетов различаются незначительно. Например, вейвлет Добеши 8 имеет центральную частоту $F_c = 0{,}6667$ Гц, и при числе отсчетов в секунду равном 16000 (частота дискретизации 16 кГц) квазичастота первого уровня декомпозиции вейвлета составит 10667,2 Гц, то есть превысит верхнюю границу речевого диапазона. Для вейвлета Добеши 16 центральная частота $F_c = 0{,}6774$, и при той же частоте дискретизации квазичастота первого уровня декомпозиции составит 10838,4 Гц. С каждым следующим уровнем разложения частота вейвлета будет уменьшаться в два раза (таблица 1). При этом частотный диапазон ниже 150 Гц не используется, так как не содержит информации, важной для задачи сегментации. Таким образом, вейвлет-коэффициенты для шести уровней декомпозиции отражают характеристики сигнала в указанном частотном диапазоне речи.

В основе выбора вейвлетного базиса лежит возможность описывать стационарный речевой сигнал со сравнительно малым числом ненулевых коэффициентов. С повышением порядка повышается гладкость функций. Таким образом, подбором порядка материнского вейвлета можно добиться наилучшего приближения.

На практике эти предположения подтвердились для случая, когда спектральные характеристики соседних

Таблица 1

Частотные диапазоны для разных уровней декомпозиции

Уровень декомпозиции	Частотный диапазон Добеши 8, Гц	Частотный диапазон Добеши16, Гц
уровень 1	5333–10667	5419–10838
уровень 2	2666–5333	2709–5419
уровень 3	1333–2666	1354–2709
уровень 4	666–1333	677–1354
уровень 5	333–666	338–677
уровень 6	166–333	169–338
уровень 7	83–166	84–169

фонем достаточно хорошо отличаются (например, сочетание «л» – «а», «и» – «о» и т.п.). Если же форма речевого тракта при переходе от фонемы к фонеме изменяется медленно, то увеличение коэффициентов детализации проявляется, как правило, только на одном уровне, который заранее неизвестен и зависит, в первую очередь, от длины сигнала и порядка вейвлета.

Решением этой проблемы является выбор адекватных вейвлетных базисов для каждого класса фонем. Выбор наиболее подходящего базиса сводится к выбору того из них, для которого количество ненулевых коэффициентов при разложении фонем данного класса минимально.

Данный мультивейвлетный подход подразумевает использование нескольких вейвлетных базисов для поиска межфонемных переходов в каждом из них с последующим объединением результатов. В качестве возможных критериев для выбора конкретного вейвлета можно также предложить свойство регулярности и число вейвлет-коэффициентов, превышающих некоторое пороговое значение. Часто для выбора используется так называемый функционал информационной ценности. В частности, рассматривается энтропийный критерий вероятностного распределения вейвлет-коэффициентов. Энтропия E функции $S(t)$ по отношению к вейвлет-базису отражает число существенных членов в разложении (1) и определяется как:

$$E = \left(-\sum_{j=1}^{n} \sum_{k=1}^{\frac{N}{2^j}-1} \left| d_{j,k} \right|^2 \log \left| d_{j,k} \right|^2 \right). \quad (3)$$

Для анализа функции $S(t)$ выбирают базис, приводящий к наименьшей энтропии (3) при минимальном числе коэффициентов детализации [11].

Длина фиксированного интервала во временной области, на котором рассчитываются признаки речевого сигнала, должна быть меньше времени звучания фонемы. В русском языке длительности фонем изменяются по разным литературным источникам в пределах 10–250 мс [12], 50–250 мс [13]. Значение длины сегмента должно обеспечить вычисление признаков речевого сигнала. Нижняя граница анализируемого частотного диапазона равна 166 Гц, в выделенный сегмент должен укладываться, по крайней мере, один период данной частотной составляющей, который равен 6 мс. Выбор величины интервала анализа является темой дополнительного исследования, выходящего за рамки данной статьи, поэтому на основании обзора литературы длина интервала, удовлетворяющая почти всем требованиям, выбирается равной 30 мс.

Предлагаемый алгоритм обработки речевого сигнала можно разбить на 3 этапа: предобработка, выделение границ, постобработка.

Этап предобработки заключается в фильтрации и нормализации сигнала.

В основе этапа выделения границ лежит высказанное выше утверждение, что при разложении сигнала с помощью вейвлет-преобразования межфонемные переходы сопровождаются быстрыми, но небольшими изменениями энергии. Рассмотрим этот этап более подробно.

– Сигнал разбивается на фреймы по 256 отсчетов с перекрытием 50%.

– Каждый фрейм накрывается окном Хэмминга для устранения краевых эффектов. Выбор типа окна является задачей отдельного исследования и не рассматривается в рамках данной статьи.

– Полученный фрейм раскладывается на 6 уровней декомпозиции с помощью кратномасштабного вейвлет-преобразования. Частотный диапазон ниже 125 Гц не используется, т.к. не содержит информации, важной для задачи сегментации. Это обусловлено природой человеческой речи, охватывающей интервал 150–4000 Гц. Так как при кратномасштабном анализе центральная частота и частотная полоса каждого следующего уровня в 2 раза больше, чем у предыдущего, то достаточно шести уровней декомпозиции.

– Для каждого из полученных уровней декомпозиции рассчитывается энергия сигнала, как сумма квадратов значений коэффициентов детализации, и записывается в результирующий массив:

$$E_j(i) = \sum_{k=1}^{\frac{N}{2^j}-1} d_{j,k}^2 , \quad (4)$$

где $E_j(i)$ – значение энергии в i-м фрейме; $d_{j,k}$ – коэффициенты детализации.

– После того как обработаны все фреймы, необходимо разбить полученный массив на пачки по 3 фрейма для того, чтобы несколько сгладить полученные значения, иначе за счет колебаний энергии от фрейма к фрейму появляются ложные границы.

– В результате имеем двумерный массив значений энергий для каждого уровня декомпозиции. Для получения информации о скорости изменения энергии $R_j(i)$ осуществляем дифференцирование. Итогом данного этапа является массив значений энергий и массив скоростей ее изменения.

– Для каждого уровня декомпозиции находятся фреймы, где абсолютная разность между энергией (4) и скоростью её изменения минимальна (меньше заданного порога), но при этом энергия больше некоторого шумового порога:

$$div_{opt} > \left| |R_j(i)| - E_j(i) \right|, \; E_j(i) > E_{\min} ,$$

где div_{opt} – пороговое значение, $R_j(i)$ – скорость изменения энергии во фрейме; $E_{\min}$ – минимальное значение энергии, не являющейся шумом. Величина пороговых значений взята: $div_{opt} = 0{,}03$, $E_{\min} = 0{,}005$ [13].

– Результатом данного этапа является двумерный массив, содержащий вектор сегментированных границ фонем для каждого уровня декомпозиции.

Этап постобработки основан на двух допущениях:

– минимальная длительность фонемы составляет 30 мс;

– алгоритм может определить границу сигнала сразу на нескольких уровнях декомпозиции, а может определить лишь на одном.

Описанный алгоритм основан на концепции «сильных» и «слабых» границ. Те границы, что найдены на нескольких уровнях, являются «сильными» и вероятность нахождения границы фонемы в таком случае высока. Границы, найденные лишь на одном–двух уровнях, являются «слабыми», их достоверность маловероятна. Данная концепция в некоторых случаях помогает предотвратить маскирование более достоверной границы ложной. Такая ситуация может произойти из-за допущения о минимальной длительности фонемы. Так, в ходе экспериментов, были получены результаты, при которых были найдены подряд две границы. Первая была «слабой», а вторая «сильной», но из-за минимальной длительности фонемы вторая, «сильная», граница не попадала в итоговую выборку, что в конечном итоге повлияло на окончательный результат сегментации, т.е. на точность. Зная о «силах» границ можно не просто проверять расстояние между границами, но и выявить, какую границу предпочтительнее вывести – ту, что раньше, но «слабее», или ту, что должна быть подавлена, но имеет большую «силу».

Экспериментальные результаты сегментации речевого сигнала

Описанный алгоритм реализован в программном пакете MatLab, в частности, с помощью пакета Wavelet Toolbox. Исследование точности сегментации с применением различных вейвлетных базисов показало, что точность сегментации с применением вейвлетного базиса Добеши 16 по критерию (3) выше, поэтому дальнейшие результаты приведены для этого базиса.

Рассмотрим результаты работы алгоритма на некоторых примерах. На рис. 1 изображены уровнеграммы энергий и их производных для каждого уровня декомпозициипри сегментации слова «невесомость» [14]. На каждом уровне вертикальными линиями отображены сегментированные границы.

На рис. 2 показан итог работы алгоритма с использованием постобработки и без неё. Уровнеграммы отображают нормированный по уровню сигнал. Полученные сегментированные границы различаются толщиной и типом линии в зависимости от числа уровней декомпозиции: 1 уровень – пунктирные тонкие; 2 уровня – штрихпунктирные тонкие; 3–4 уровня – штрихованные утолщенные; более 5 уровней – сплошные утолщенные.

Сравним результаты работы алгоритма с постобработкой и без неё. Без использования постобработки отображается большое количество ложных границ (рис. 2), показывающих, что границы фонем выделяются неоднозначно. Кластеры границ образуются за счет того, что алгоритм выделяет границы на нескольких уровнях детализации, соответствующих различным частотным промежуткам сигнала. Такой эффект образуется вследствие того, что межфонемные переходы происходят не мгновенно, а с небольшими задерж-

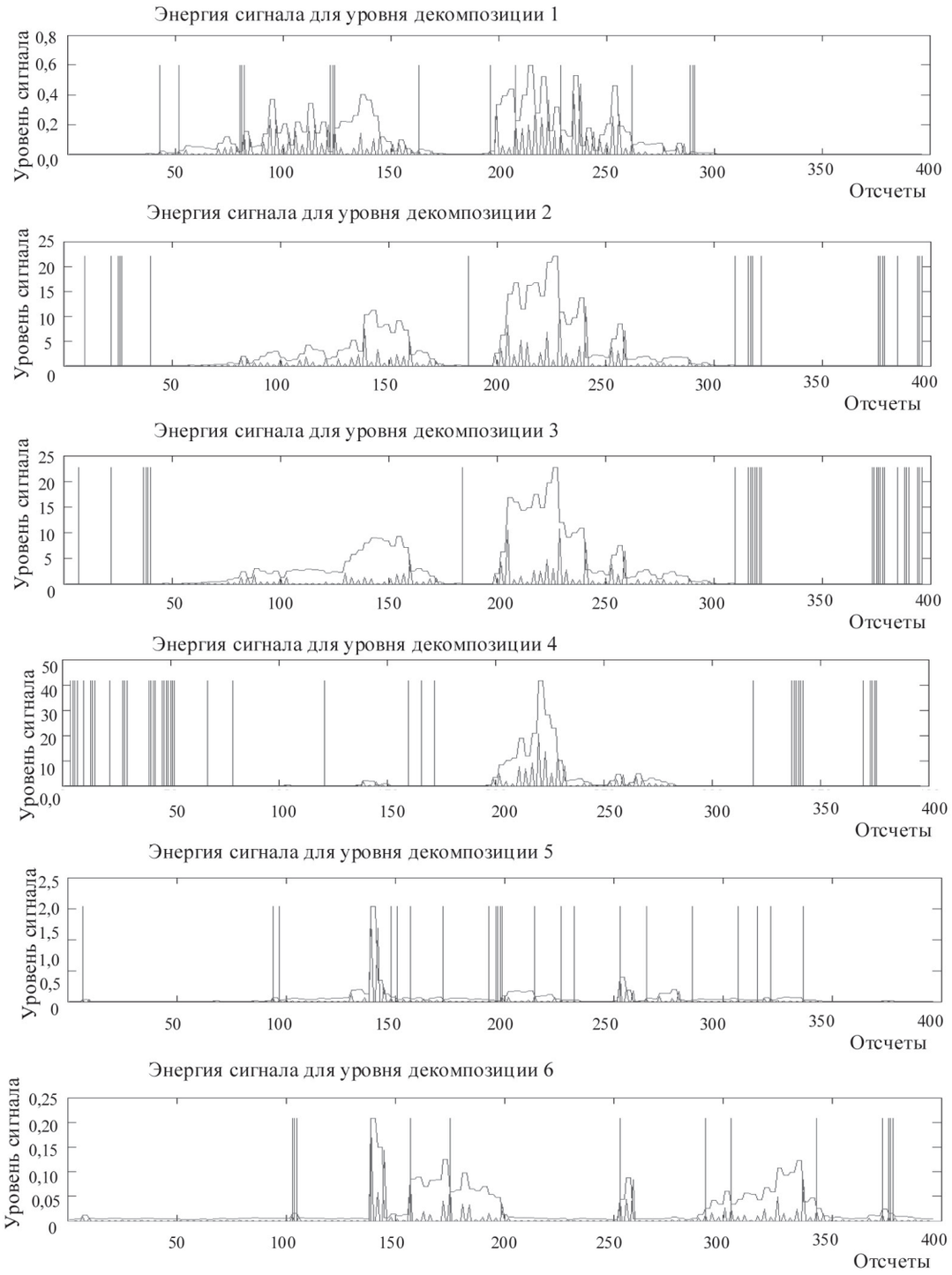

Рис. 1. Графики зависимости энергии и скорости ее изменения для слова «невесомость»

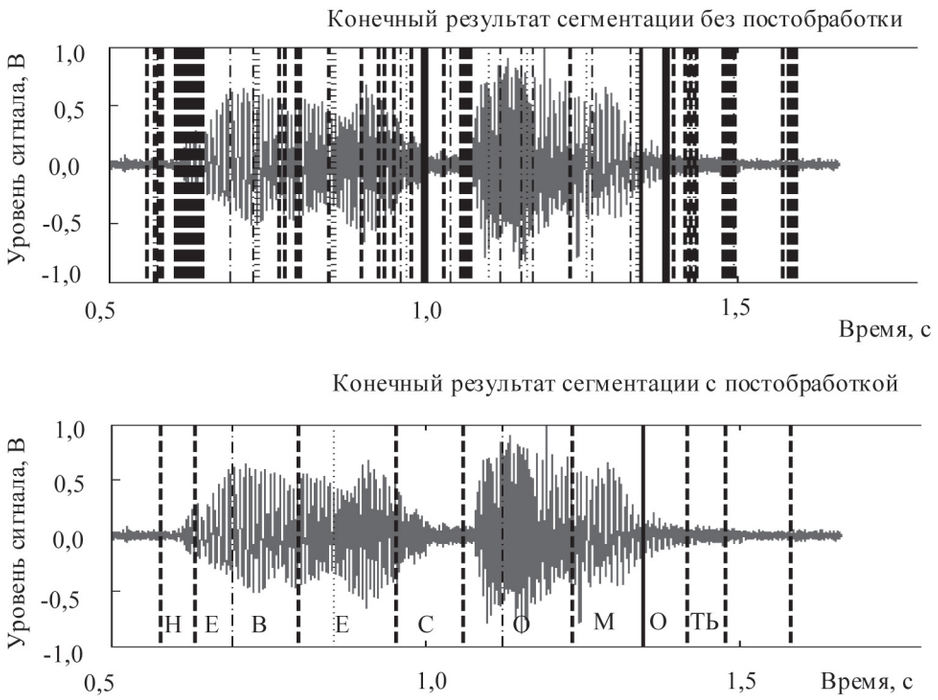

Рис. 2. Результат обработки слова «невесомость»

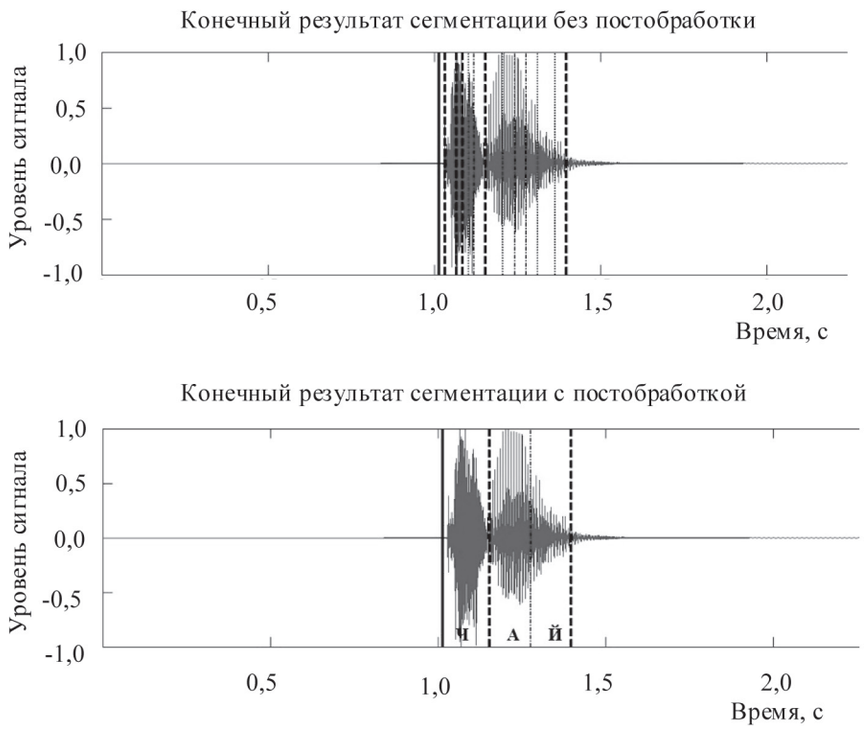

Рис. 3. Результат обработки слова «чай»

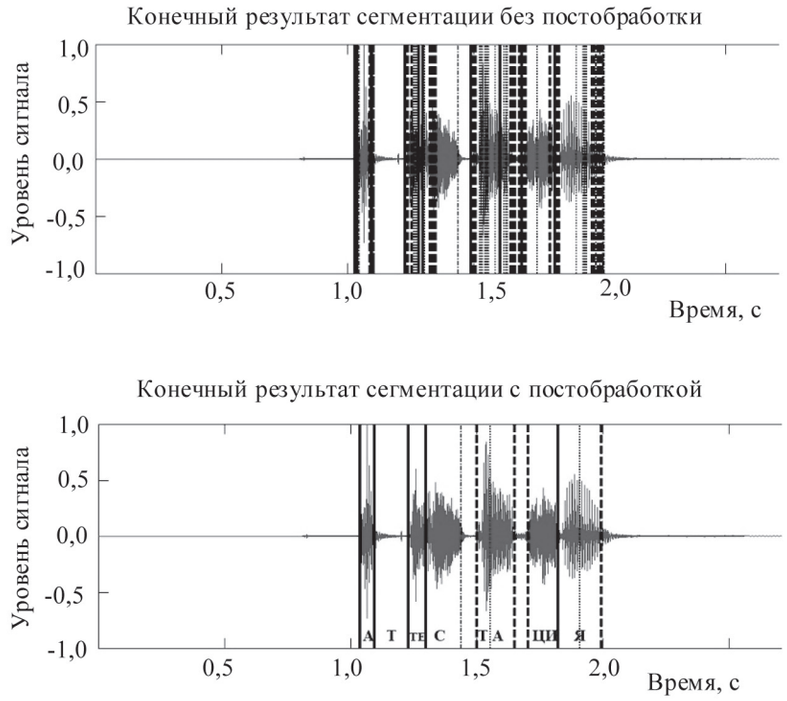

Рис. 4. Результат обработки слова «аттестация»

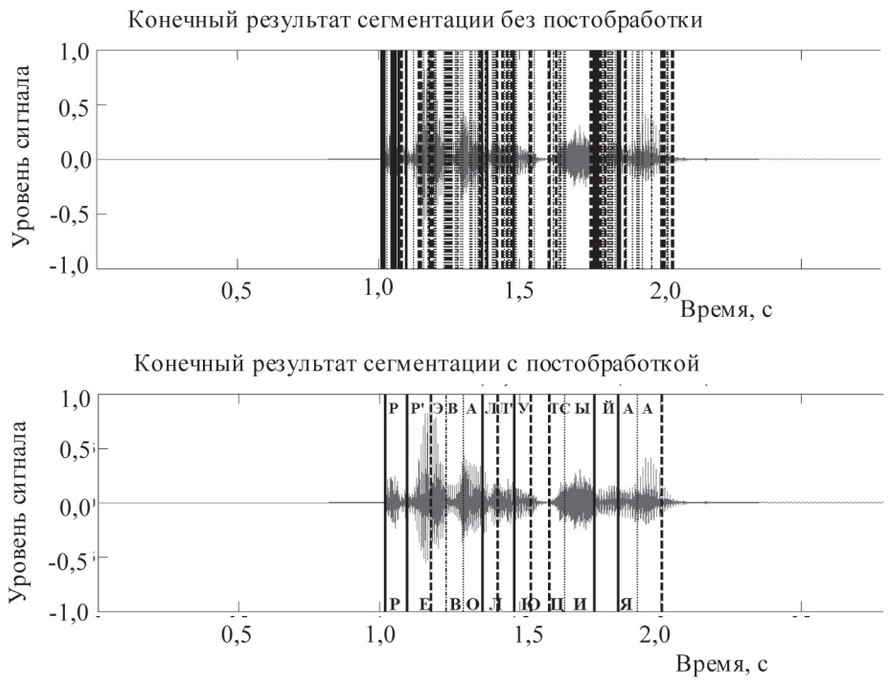

Рис. 5. Результат обработки слова «революция»

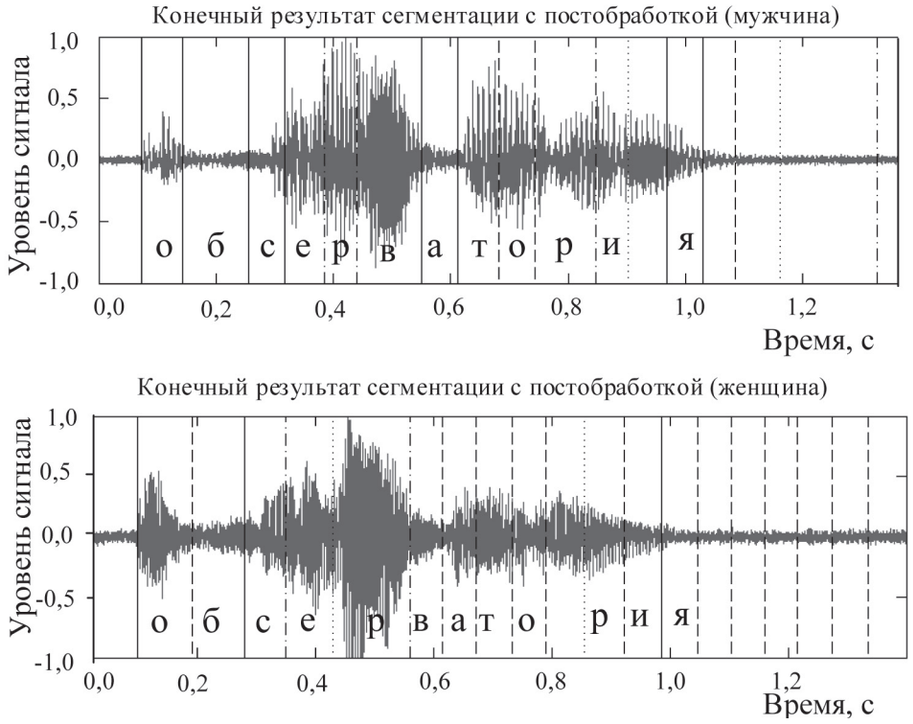

Рис. 6. Результат обработки слова «обсерватория»

ками, т.е. сигнал изменяется не одновременно на всех частотных диапазонах, а поочередно (в зависимости от содержащихся в слове фонем).

Интерес представляет сегментация слов, содержащих фонетические дифтонги, т.е. сочетание в одном слоге двух гласных – слогового и неслогового, например, звуковое сочетание «ай», в слове «чай» (рис.3).

В результате работы алгоритма без постобработки было выделено большое число ложных границ (рис. 3). Постобработка позволила убрать ложные границы и выделить 3 фонемы: «ч», «а», «й». То есть алгоритм точно определил границы фонем в сложном слове, содержащем дифтонг. Аналогичный результат был получен при сегментации этого же слова, произнесенного другим человеком.

Стоит отметить, что слово «чай» помимо фонетического дифтонга включает в себя аффрикату «ч», представляющую собой слитное сочетание смычного согласного с фрикативным [1]. Несмотря на то, что аффриката «ч» состоит из «ть» + «щ», данная фонема была выделена как цельная лингвистическая единица без разбиения на составляющие. При необходимости выделения составляющих сложных звуков, можно изменить значения коэффициентов, использующихся при постобработке.

Более сложным для задачи сегментации является слово «аттестация» (транскрипция слова – [ат′:эстацый′а]), так как содержит:

– долгую согласную «т». Долгие согласные в русском языке должны рассматриваться как сочетания двух кратких: «аттестация – ат-тестация»;

– сибилянт «ц» – согласный звук, при произношении которого поток воздуха стремительно проходит между зубами [1];

– сложное окончание «ия».

Проанализируем полученные результаты, приведенные на рис. 4.

· Долгая согласная «т» была распознана как 2 коротких звука, но алгоритм не установил границы между звуками «т'» и «э», восприняв их как единую фо-нему. Однако, рассмотрев результат работы алгоритма, без постобработки можно отметить наличие «сильной» границы между данными фонемами, которая была потеряна после применения правила о минимальной длительности фонемы. Это является одним из минусов предлагаемого алгоритма постобработки.

· Сибилянт «ц» был распознан не как отдельный звук, а как сочетание «ци». Видно, что в данном случае постобработка так же привела к потере необходимой границы.

· Точно найдены границы в сложном окончании «ия» между буквами «и» и «я». Стоит отметить, что алгоритм выделил составляющие буквы «я», разбив её на звуки «й» и «а».

Рассмотрим сегментацию слова «революция» (рис. 5), транскрипция которого не всегда точно совпадает с его произношением, что показано на рис. 5б. Фонетическая транскрипция данного слова – [р'ивал'уций'а], но фактически транскрипция изменяется в зависимости от того, как произносится это слово. В данном примере были получены следующие звуки: [р'эвал'утсый'а]. Фонетическая транскрипция не учитывает характера звуков, например, в слове «революция» присутствует аффрикативный звук «ц», состоящий из звуков «т'» и «с». Алгоритм выделил все необходимые границы, однако звук «р» был разделен на «р» и «р'», т.е. был воспринят как долгая согласная. Это не является ошибкой алгоритма, а является следствием произношения конкретного человека. Плюсом алгоритма является точная сегментация звука «ц», так как он не был разделен на «т» и «с», что облегчает дальнейшую обработку в процессе распознавания.

На примере сегментации слова «обсерватория» рассмотрим зависимость точности работы алгоритма от того, кто произнес это слово – мужчина или женщина. На рис. 6 изображены результаты работы алгоритма сегментации речевого сигнала с постобработкой для этого случая. В результате анализа установлено, что на точность работы алгоритма практически не влияет является ли говорящий мужчиной или женщиной, однако большое влияние оказывает скорость произношения и артикуляция говорящего. В частности, из-за разницы в скорости произношения слова мужчиной границы фонем «о'» и «р'» были точно сегментированы, в отличие от результата сегментации слова, произнесенного женщиной. Так как слово было произнесено быстрее, алгоритм неточно определил смену фонем и сместил границу.

Большинство исследователей для определения точности работы своих алгоритмов используют речевые базы с закрытым доступом. Каждый речевой сигнал, содержащийся в подобных базах, сегментируется вручную опытным лингвистом. Автоматически выделенные границы сравнивают с поставленными вручную, и рассчитывается приемлемая погрешность автоматической расстановки. Для определения точности распознавания границ по реализованному алгоритму были использованы 23 слова, содержащих разное количество фонем (от 3 до 14), произнесенные людьми разных возрастов и пола. В результате из 115 слов точно было сегментировано 87, что соответствует точности работы алгоритма – 76%. В результате эксперимента было выяснено, что пол и возраст говорящего не оказывают влияния на точность работы алгоритма, незначительные колебания в точности сегментации возникают из-за разницы в скорости произношения и артикуляции.

Заключение

Преимуществами описанного алгоритма сегментации речевого сигнала на базе вейвлетов Добеши с применением постобработки являются:

· Высокая точность сегментирования границ фонем. Алгоритм продемонстрировал хорошие результаты при сегментации долгих согласных, аффрикативных согласных и сибилянтов.

· Алгоритм показал слабую зависимость от пола и возраста говорящего.

· Наличие постобработки повышает точность сегментирования, уменьшая вероятность нахождения ложных границ. Без постобработки алгоритм находит границы в аффрикативных согласных, которые усложняют дальнейший процесс распознавания.

К недостаткам алгоритма следует отнести:

· Пороги выбираются эмпирическим путем, т.е. для повышения точности распознавания необходимо вручную корректировать коэффициенты, подстраивая алгоритм под те или иные особенности речевого сигнала, такие как артикуляция и скорость произношения говорящего.

· В некоторых случаях постобработка стирает нужные границы.

Таким образом, при сегментировании речевого сигнала с помощью алгоритма, построенного на базе вейвлет-преобразования, с использованием постобработки можно довольно точно определить границы фонем. Стоит отметить, что в дальнейшем при осуществлении незначительных изменений, данный алгоритм может использоваться в качестве алгоритма VAD (VoiceActivityDetection), т.к. позволяет определять не только межфонемные переходы, но и переходы тишина-речь.

Литература

1. Скрипник, Я. Н. Фонетика современного русского языка / Я.Н. Скрипник, Т.М. Смоленская; под ред. Я.Н. Скрипник. – Ставрополь: Изд–во СГПИ, 2010. – 148 с.

2. Separation of Voiced and Unvoiced Using Zero Crossing Rate and Energy of the Speech Signal / R.G. Bachu [et al.] // American Society for Engineering Education (ASEE) Zone Conference Proceedings, 2008. – P. 1–7.

3. Moattar, M. H. A simple but efficient real-time voice activity detection algorithm / M.H. Moattar, M.M. Homayoonpoor // 17th European Signal Processing Conference (EUSIPCO 2009). Glasgow, Scotland, 2009. – P. 2549–2553.

4. Mattias, N. Entropy and Speech / N. Mattias // Sound and Image Processing Laboratory School of Electrical Engineering KTH (Royal Institute of Technology). Stockholm, 2006. – 54 p.

5. Jancovic, P. Estimation of Voicing-Character of Speech Spectra Based on Spectral Shape / P. Jancovic, M. Kokuer // IEEE Signal Processing Letters. – 2006. – Vol. 14, No. 1. – P. 66–69.

6. Демидов, А. Е. Использование меры спектрального перехода для сегментации речевого сигнала / А.Е. Демидов, В.М. Смирнов // Материалы 66 Международной студенческой научной конференции ГУАП, СПб, 2013. – С. 119–122.

7. Ahmadi, S. Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm / S. Ahmadi, A.S.

Spanias // IEEE Transactions on Speech and Audio Processing. – 2002. – Vol. 7, No. 3. – P. 333–338.

8. Robust Voiced/Unvoiced Classification Using Novel Features and Gaussian Mixture Model / J.K. Shah [et al.] // IEEE International Conference on Acoustics, Speech, and Signal Processing. – 2004. – P. 17–21.

9. Juang, C. F. Speech detection in noisy environments by wavelet energy-based recurrent neural fuzzy network / C.F. Juang, C.C. Nan // Experts Systems with Applications. – 2009. – Vol. 36 (1). – P. 321–332.

10. Сорокин, В. Н. Сегментация и распознавание гласных / В.Н. Сорокин, А.И. Цыплихин // Информационные процессы. – 2004. – Т. 4, № 2. – С. 202–220.

11. Желтов, П. В. Применение быстрого непрерывного вейвлет-преобразования для исследования акустических сигналов / П.В. Желтов, В.И. Семенов // Вестник Чувашского университета. Естественные и технические науки. – 2010. – № 3. – С. 309–312.

12. Медведев, М. С. Фонемная сегментация речевого сигнала с использованием вейвлет-преобразования / М.С. Медведев // Материалы Всероссийской конференции молодых ученых по математическому моделированию и информационным технологиям с участием иностранных ученых, 2004. – С. 110–124.

13. Вишнякова, О. А. Автоматическая сегментация речевого сигнала на базе дискретного вейвлет-преобразования / О.А. Вишнякова, Д.Н. Лавров // Математические структуры и моделирование. – 2011. – №. 23. – С. 43–48.

14. Smirnov, V. M. Application of wavelet transform for speech signal segmentation / V.M. Smirnov, V.N. Filatov // XXII International Scientific Conference Wave Electronics and ITS Application in Information and Telecommunication Systems WECONF-2019.