УДК 004.522

В.О. Верходанова

АЛГОРИТМЫ И ПРОГРАММНЫЕ СРЕДСТВА АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ РЕЧЕВЫХ СБОЕВ В ЗВУКОВОМ СИГНАЛЕ

Верходанова В.О. **Алгоритмы и программные средства автоматического определения речевых сбоев в звуковом сигнале**

Аннотация. При автоматической обработке спонтанной речи возникает ряд трудностей, таких как вариативность речи или присутствие речевых сбоев различной природы. В статье рассматриваются различные виды речевых сбоев и причины их возникновения, а также представлен алгоритм их автоматического определения, основанный на анализе акустических параметров. Для выделения звонких хезитационных явлений использовался кросскорреляционный метод, а для выделения глухих хезитационных явлений – метод полосовой спектральной фильтрации. Эксперименты проводились на специально собранном корпусе спонтанной русской речи, состоящем из диалогов по описанию маршрута по карте и нахождению общего свободного времени по расписанию. Проведенные эксперименты показали, что звонкие хезитационные явления выделяются с точностью 80%, глухие хезитационные явления и дыхание - с точностью 66%.

Ключевые слова: речевые сбои, речевой корпус, автоматическая обработка речи, автоматическое распознавание речи.

Verkhodanova V.O. **Algorithms and Software for Automatic Detection of Speech Disfluencies in an Audio Signal**

Abstract. During automatic speech processing a number of problems appear, and among them are such as speech variation and different kinds of speech disfluences. In this article different types of speech disfluencies and their causes are presented, as well as the algorithm for their automatic detection based on the analysis of acoustical parameters. The method of cross-correlation was used to detect voiced hesitation phenomena and a method of band-filtering was used to detect unvoiced hesitation phenomena and artefacts. The experiments were performed on a specially collected corpus of spontaneous Russian map-task and appointment-task dialogs. Experiments showed that voiced hesitation phenomena are detected with 80% accuracy and devoiced hesitation phenomena and artefacts – with 66% accuracy.

Keywords: speech disfluencies, speech corpus, automatic speech processing, automatic speech recognition.

1. Введение. Спонтанная речь характеризуется, прежде всего, неподготовленностью формы. Ей свойственны ошибки, наличие речевых сбоев (например, хезитаций), большая вариативность, небрежность, неполнота, недостаточное внимание к внешней форме, что невозможно в письменной речи, однако естественно для устной коммуникации. Изучение речевых сбоев началось еще в 50х годах XX-го века в рамках психологии, лингвистики и физиологии. Сейчас речевые сбои активно изучаются в различных областях науки. Так, в области социолингвистического анализа бытового диалога речевыми сбоями занимался

Б.Фокс [13]. Одним из первых психолингвистические исследования спонтанной речи проводил Х.Кларк [11], экспериментально к исследованию речевых сбоев в психолингвистике подошел М.Корли [12].

С развитием технологического прогресса все больше внимания стало уделяться автоматической обработке речевого сигнала, поскольку голосовое взаимодействие с мобильными устройствами становится все более востребованным. Примером служит вопросно-ответная система Siri (Speech Interpretation and Recognition Interface), где развиваются такие технические области, как диалоговый интерфейс, понимание контекста и сервисная делегация. Другим примером является Google Voice Search, сервис голосового поиска от компании Google, который, помимо английского, китайского, корейского, японского, французского, немецкого, итальянского, польского и чешского, в настоящее время поддерживает также и русский язык.

Несмотря на актуальность вопроса и многосторонние исследования речевых сбоев, общепринятая терминология в этой области пока не сложилась. Для описания этих явлений существуют различные варианты терминов. В англоязычной литературе можно встретить такие термины, как «non-fluency», «disfluency», «discontinuity», «flustered speech», «speech disturbance», «speech management», «own communication management», «hesitation», «turnholding devices» и др. [29]. В русскоязычной литературе эти явления иногда рассматривают в рамках фонационных паралингвистических явлений, и можно встретить термины «внеязыковые элементы речи», «речевые сбои» [1, 3, 6]. Отдельно стоит рассматривать артефакты речи — неречевые физиологические явления, такие как кашель, вздох, смех и т. п.

Определение подобных речевых сбоев и артефактов актуально как для обработки звучащей речи, так и ее транскрипций, поскольку они являются источником ошибок при распознавании, что приводит к ошибкам в транскрипциях. Таким образом, разработка алгоритмов автоматической обработки речевых сбоев стала в настоящее время чрезвычайно актуальной задачей для всех языков. В данной работе исследуются такие речевые сбои, как заполненные паузы хезитации и хезитационные удлинения.

2. Типы речевых сбоев, причины их появления и способы автоматического определения. Возникновение сбоев в речи может быть вызвано внешними воздействиями и сбоями в планировании речевого акта [6]. Эти явления имеют разную природу, и среди них выделяют заполненные паузы хезитации, самокоррекции (или самоисправления), оговорки. Учитывая различные причины возникновения

речевых сбоев, можно ввести следующую классификацию, как показано на рисунке 1.

Паузы хезитации - паузы колебания, хезитационные паузы - представляют собой перерыв в фонации при порождении высказывания, часто заполняются некоторыми звуками. Обычно такие паузы представляют собой семантические лакуны и свидетельствуют о том, что говорящему требуется дополнительное время на формулирование следующего за текущим фрагмента высказывания [5, 7, 11]. Существуют разные типы заполнения пауз хезитации [6]:

1. Абсолютная пауза.
2. Удлинение отдельных звуков в словах.
3. Словоподобные, «долексические» заполнения.
4. Вспомогательные элементы дискурса (слова и словосочетания).

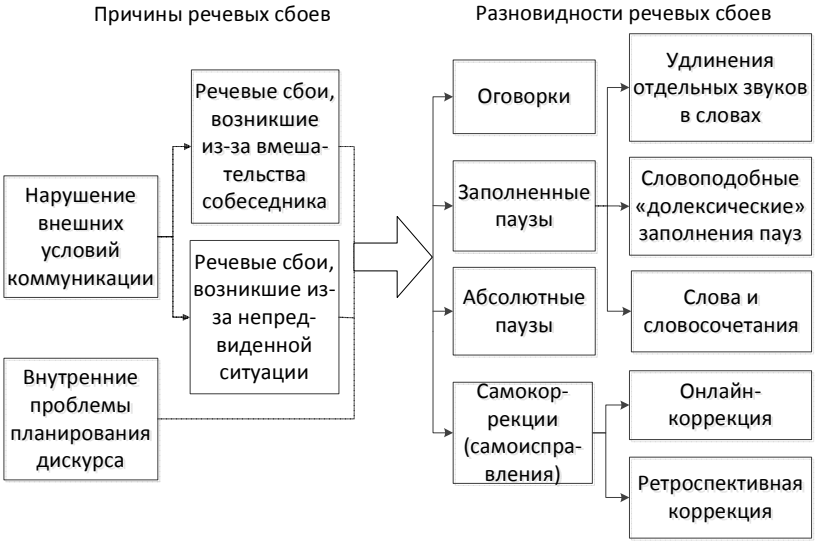

Рис.1. Классификация причин и видов речевых сбоев.

С.Б. Степанова выделила двенадцать способов заполнения хезитаций, различающихся по месту появления в синтагме (начало, середина, конец): абсолютная пауза, речеподобный звук, повтор, заполнение словами-паразитами, оговорки и запинки, а также хезитационные удлинения звуков («затяжки») [7].

В своей работе С.Б. Степанова также привела данные зарубежных исследователей, изучавших акустические характеристики заполненных пауз хезитации. По ее мнению данные несколько противоречат друг другу [7]. Так, исследовавший английскую спонтанную речь О'Шонесси пришел к выводу, что спектральная картинка заполненных пауз соответствует картинке звука шва, отмечая, однако, что этот звук может реализоваться по-разному в разных языках [26]. Исследователи немецкой спонтанной речи Петцолд и Симпсон, заключили, что заполнение пауз хезитации происходит звуками, не входящими в фонетическую систему языка и представляющими собой центральный гласный [27]. На материале шведского языка, В. Левельт обнаружил заполнение хезитационных пауз передним открытым гласным [21]. А. Гианини, исследуя итальянскую речь, пришел к выводу, что по качеству заполнения хезитационных гласных можно отличить говорящих из разных частей Италии, поскольку это зависит от региональной фонологической системы дикторов (дикторы из Неаполя заполняли паузы центральными гласными, дикторы из Пизы и Рима — передними закрытыми или открытыми гласными) [14].

Как пишет в своей работе Э. Шриберг [30], паузы хезитации, заполненные долексическими элементами рассматривались по-разному. Были работы, которые классифицировали их вместе с артефактами (смехом, кашлем, вздохом). Другие авторы считали их с языковыми элементами, объединяя с сочинительными союзами [10] или с дискурсивными маркерами, такими как: “ты знаешь” или “ну вот” [8].

В зарубежной литературе принято говорить о временных характеристиках сбоев. В работе Э. Шриберг [30] используются следующие термины:

- reparandum (кратко RM) – репарандум, участок звукового сигнала, соответствующий всему удаленному отрезку речи;

- interruption point (IP) – начало речевого отрезка, соответствующее «моменту прерывания» чистой речи и возникновению речевого сбоя;

- interregnum (IM), (у других авторов «редактирующая фаза»[20] или «интервал сбоя» [25]) – длительность речевого сбоя. Этот термин используется для обозначения временного отрезка от RM до начала исправления. При этом он может не содержать редактирующего элемента. Например, незаполненная пауза может быть использована говорящим для перепланировки высказывания без редактирования;

- repair (RR) – исправление, это участок речи, который соответствует материалу репарандума.

Если же речевой сбой является хезитацией, то она попадает на участок IM. На рисунке 2 показан пример коррекции фразы.

Рис.2. Пример коррекции фразы.

В этом примере *из Москвы* - репарандум, э-э — интеррегнум, *из Петербурга* — исправление, а момент прерывания попадает между *Москвы* и э-э.

В настоящее время для исследования речевых сбоев используются корпусы спонтанной речи с многоуровневой аннотацией. Для того чтобы в корпусе помимо такой информации, как фонемы, слова, синтагмы, дополнительно отмечать речевые сбои, используется Rich Transcription – транскрипция, в которой учитываются границы предложений, слова-заполнители, речевые сбои и др. [22]. Примером служит корпус SWITCHBOARD [33], который представляет собой более 240 часов записи спонтанных телефонных разговоров более 500 дикторов обоих полов. Речь полностью затранскрибирована, и транскрипции проверены автоматически и дважды вручную. Неречевые данные отмечены в квадратных скобках. Всего размеченных типов неречевых данных 78, среди которых были отмечены вздох, кашель, зевание, мяуканье, звон посуды и др.

Также стоит упомянуть аннотированный корпус русской речи, разработанный на кафедре фонетики СПбГУ. Этот корпус включает в себя речь 4 мужчин и 4 женщин и учитывает различные произносительные стили. Аннотация производилась на 6-ти уровнях, где отмечалась вся фонетическая и просодическая информация о записанной речи [31].

Для исследования речи с патологиями существует архив речи с заиканиями «Archive of Stuttered Speech» [15], созданный Университетским колледжем Лондона. Архив состоит в основном из записей заикающихся детей школьного возраста. Записи доступны в разных форматах с транскрипциями и расшифровками.

Методы обработки речевых сбоев можно разделить по признаку описания их с помощью акустических моделей или с помощью комбинированных моделей (языковые + акустические). Чаще всего используются только акустические модели речевых сбоев для их последующего применения в системах автоматического распознавания речи.

Множество работ посвящены моделированию речевых сбоев в рамках создания систем автоматического распознавания речи [17, 23, 24]. Помимо этого существуют подходы, направленные на повышение качества распознавания спонтанной речи за счет предварительного выявления речевых сбоев и их устранения из звукового сигнала на этапе цифровой обработки [16], или устранения сбоев с использованием транскрипций речи [23, 32]. Методы параметрической обработки речевых сбоев опираются только на акустические параметры [16, 24], тогда как комбинированные модели используют дополнительно языковое моделирование [19, 23, 32].

В работе [28] показано, что улучшение качества распознавания речи возможно за счет выделения в сигнале свойств сложных речевых сбоев, которые представляют собой смесь хезитаций и самокоррекций. Авторы предлагают рассматривать спектральную область на основе мел-кепстральных коэффициентов MFCC (Mel-Frequency Cepstral Coefficients) одновременно с символьными последовательностями, описывающими сбои. На небольшой выборке из корпуса UCLASS [15] по речи с заиканиями авторы продемонстрировали улучшение качества распознавания при использовании разработанного алгоритма с 50.2% до 97.6% по сравнению с MFCC.

Для применения языковой модели дополнительно необходим большой корпус транскрипций, по которому происходит обучение модели, в то время как параметрическая обработка не требует транскрибирования сигнала. Поэтому в данной работе было решено исследовать параметрические методы для цифровой обработки аудиосигнала и определения границ речевых сбоев.

3. Описание собранного корпуса спонтанной русской речи. Неофициальность отношений между собеседниками, их непосредственное участие в разговоре, устная форма, опора на внеязыковую ситуацию, использование в основном жанра диалога являются основными условиями непринужденного неподготовленного непосредственного общения [2]. Для соблюдения всех названных аспектов была разработана методика, позволяющая записать в контролируемых условиях корпус, максимально приближенный к разговорной речи. Участники записи выполняли два задания: они должны были 1) описать

маршрут по карте от одного пункта до другого (т.н. map-task), 2) определить время встречи на основе своих личных расписаний (т.н. appointment-task).

Подобный подход к записи корпуса использовался еще в 90-х годах. В Эдинбурге и Глазго был записан корпус HCRC, состоящий только из диалогов-описаний маршрутов по карте. В него входит 128 диалогов, 18 часов, в записи которых принимали участие 64 человека, 32 мужчины и 32 женщины. Карты и названия пунктов были разработаны таким образом, чтобы исследовать различные фонологические процессы в речи. Транскрипции содержали помимо слов пометки для речевых сбоев [9]. Диалоги по нахождению общего свободного времени на основе расписаний составляют половину другого корпуса — корпуса немецкой речи Kiel, состоящего из четырех часов чтения и из четырех часов диалогов на основе расписаний. В разметке корпуса учитывались сегментные и просодические характеристики [18].

Задача описания маршрута по карте состоит в том, что один участник записи должен объяснить другому дорогу по карте от начального пункта до конечного. Участникам выдаются карты (одна с указанным маршрутом, другая – без него), при этом говорящие не видят карт друг друга. Детальное описание карт приведено в работе [34]. Участник с картой, где обозначен маршрут, объясняет другому участнику, куда ему следует двигаться по карте. Затем участники меняются ролями, получают новые карты и диалог продолжается. При этом участников предупреждают о том, что некоторые объекты на картах могут не совпадать. В рамках данного исследования было создано несколько комплектов карт разной сложности с разными типами объектов на картах. Критерием сложности служило количество несовпадающих объектов на картах участников диалога.

Задача определения общего свободного времени на основе личных расписаний заключалась в том, что собеседники должны были найти время для: а) телефонного разговора длительностью в 15-20 минут, б) встречи длительностью 1 час. Расписания были сформированы так, чтобы не было однозначного решения задачи [34].

Поскольку содержание как карт, так и названий различается, участники вынуждены задавать вопросы, переспрашивать, перебивать и уточнять, что приводит к порождению различных речевых сбоев.

В рамках проекта РФФИ 12-06-31203 мол_а был собран корпус русской спонтанной речи, часть которого послужила материалом для данного исследования. Эта часть корпуса представляет собой 50 минут диалогической речи - 18 диалогов длительностью от 1,5 до 5 минут.

Запись проводилась в звукоизолированной комнате с использованием современных мобильных устройств (планшетов) на базе ОС Android Samsung Galaxy Tab 2, приложением Smart Voice Recorder. Записанные аудиофайлы имеют следующие параметры: частота дискретизации 16кГц, битрейт – 256 кбит/с, количество каналов – 1. Все записи были сделаны в Санкт-Петербурге в конце 2012 – начале 2013 года. В записи принимали участие 12 человек: 6 девушек, 6 юношей в возрасте от 17 до 23 лет технических и гуманитарных специальностей (по три представителя каждого пола технических и гуманитарных специальностей) с полным или на момент записи еще неоконченным высшим образованием. Дикторы были между собой знакомы или же находились в дружественных отношениях, что способствовало непринужденности и неофициальности общения. Устная форма, опора на внеязыковую ситуацию, неподготовленность речи и необходимость диалога обеспечивались заданиями. Таким образом, можно говорить о спонтанности и разговорности собранного речевого материала.

Корпус был вручную аннотирован в программе Wave Assistant. Обозначение пауз хезитаций и артефактов строилось по схеме «хезитация . заполнитель» (h.filler – hesitation.filler), «артефакт . заполнитель» (ar.filler – artifact.filler). Разметка производилась на 2-х уровнях: на одном уровне отмечались явления, встретившиеся в речи одного диктора, на втором – в речи другого. В ходе аннотации были размечены такие элементы, как заполненные паузы хезитации (например, [ə], [ɐ]), артефакты речи (например, смех, вздох), самокоррекции и фальстарты, а также слова и словосочетания, заполняющие паузы.

Всего было выделено 1042 явления речевых сбоев и артефактов, которые можно разделить на условные классы. В класс артефактов объединялись такие элементы, как вздох, смех, кашель, причмокивание; в класс хезитаций объединялись заполненные паузы типа [ə] или [ɐ]; в класс удлинений (хезитационных удлинений) объединялись хезитационные растяжки звуков; самоисправления и фальстарты объединялись в один класс самокоррекций, и отдельно был выделен класс слов-паразитов – заполнителей пауз. Самые частотные элементы классов показаны на рисунке 3.

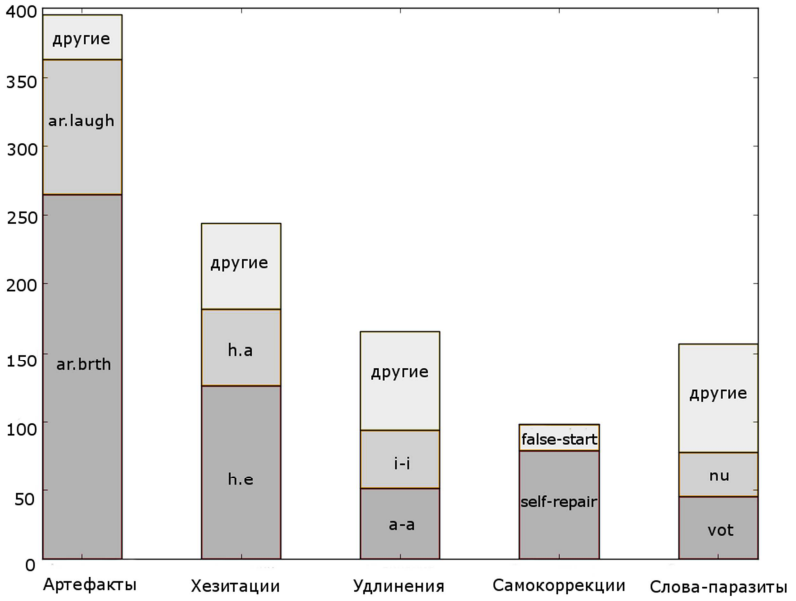

Рис.3. Распределение элементов по классам (где ar.laugh - смех, ar.brth - вздохи, h.a – хезитация [ɐ], h.e – хезитация [ə], i-i – хезитационное удлинение /i/, a-a – хезитационное удлинение /a/, false-start -фальстарт, self-repair - самокоррекция, nu – «ну», vot – «вот»)

В речи всех 12 дикторов встретилось одинаково часто дыхание (вздох), паузы хезитации [ə] и [m], самокоррекции и хезитационные удлинения гласного /i/. Для 11 говорящих кроме уже перечисленного, общими являлись хезитационное удлинение гласного /a/ и хезитация [ɐ]. Также почти всем говорящим были свойственны слова-паразиты «вот» и «ну».

4. Алгоритм автоматического определения речевых сбоев и артефактов в сигнале. В силу того, что некоторые речевые сбои мало отличаются от обычной речи и обладают коммуникативной значимостью, на данном этапе исследования было решено их не рассматривать и ограничиться заполненными паузами хезитации и хезитационными удлинениями. Одними из отличительных характеристик данных элементов являются постоянство частоты основного тона (ЧОТ), постоянное расположение формант в спектре и длительность, превышающая 150-200 мс. Надежное восприятие пауз начинается именно с этого значения, поскольку оно близко к значению средней длительности слога

[4]. Этим явлениям свойственно постоянство ЧОТ, но не все интервалы, где ЧОТ постоянна, можно отнести к этим явлениям. Например, слова «мне», «невменяемый», «налево» будет иметь длинный участок непрерывной и постоянной ЧОТ (см. рис.4).

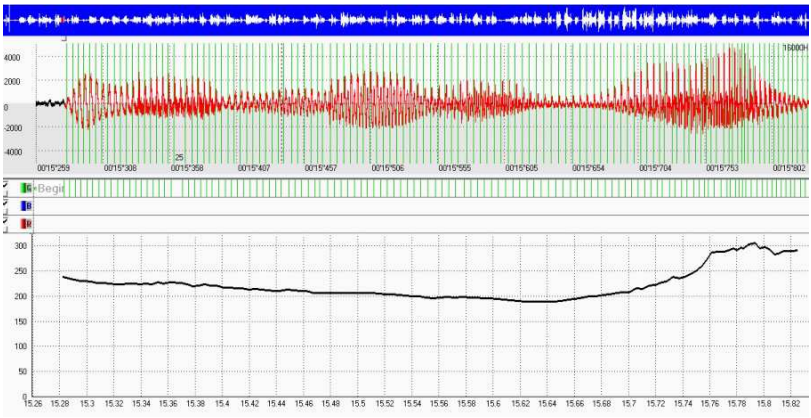

Рис.4. График частоты основного тона для слова «налево» (интервал усреднения 50мс).

Для решения этой проблемы было решено отслеживать изменения во всей структуре спектра, а не только в частоте основного тона. Анализ степени постоянства спектра сигнала производился на основе расчета кросскорреляционной функции нормированных спектров соседних сегментов речи [34]. В случае постоянства спектра значение максимума кросскорреляционной функции приближается к единице. В противном случае это значение уменьшается.

В ходе посегментного анализа (в окне длительностью 512 отсчетов с шагом 256 отсчетов) речевого сигнала предварительно производилось вычисление следующих пяти характеристик: 1) энергия сигнала *E*; 2) спектр сигнала *S*; 3) отношение сигнал/шум *SNR*; 4) отношение среднего значения спектра внутри каждой из полос, заданных для глухих щелевых согласных, к среднему значению по всему спектру *R*; 5) значение максимума кросскорреляционной функции между спектрами соседних сегментов речи *C*.

В общем виде схема работы алгоритма показана на рисунке 5. В результате параметрической обработки звукового сигнала определяются перечисленные выше пять характеристик, обеспечивающих вы-

деление наборов интервалов звонких и глухих хезитационных удлинений, а также интервалов дыхания. Определенные автоматически наборы интервалов используются при сравнении с ручной разметкой этих явлений. Поиск звонких хезитационных явлений (пауз хезитации и хезитационных удлинений) производился на основе алгоритма, показанного на рисунке 6.

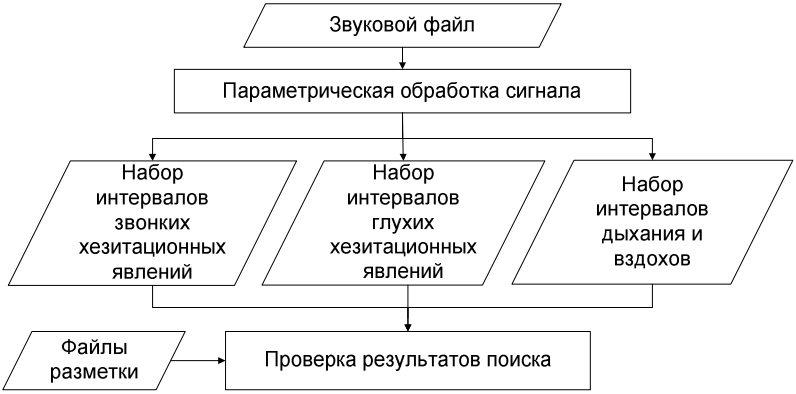

Рис. 5. Общая схема работы алгоритма автоматического определения хезитационных явлений и дыхания.

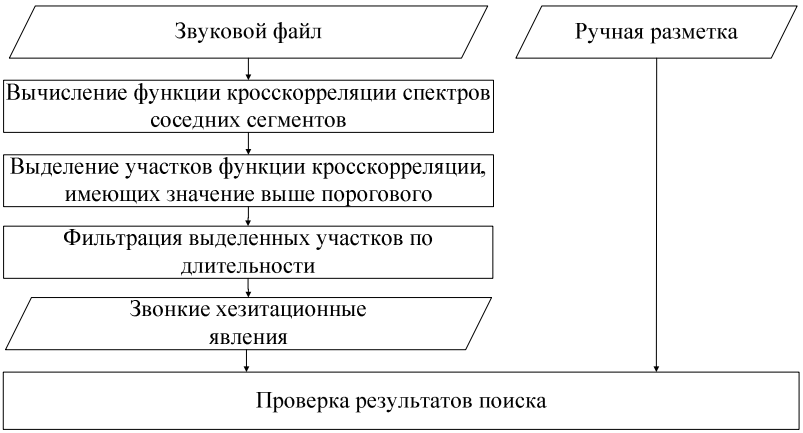

Рис. 6. Схема поиска звонких хезитационных явлений.

Вначале производился расчет кросскорреляционной функции спектров соседних сегментов, затем определялись участки, на которых значения кросскорреляционной функции близки к единице. Получившиеся интервалы соответствуют относительно продолжительным гласным и сонантам (минимальная длительность - 2 окна). Затем отбирались интервалы, длительность которых превышает заданный экспериментально порог (0,18с): такие интервалы считались соответствующими искомым удлинениям [34].

Что касается рассматриваемых в данном исследовании артефактов, были взяты наиболее распространенные из них: вздохи и громкое дыхание. Поиск дыхания, удлиненных глухих щелевых согласных производился на основе алгоритма, показанного на рисунке 7.

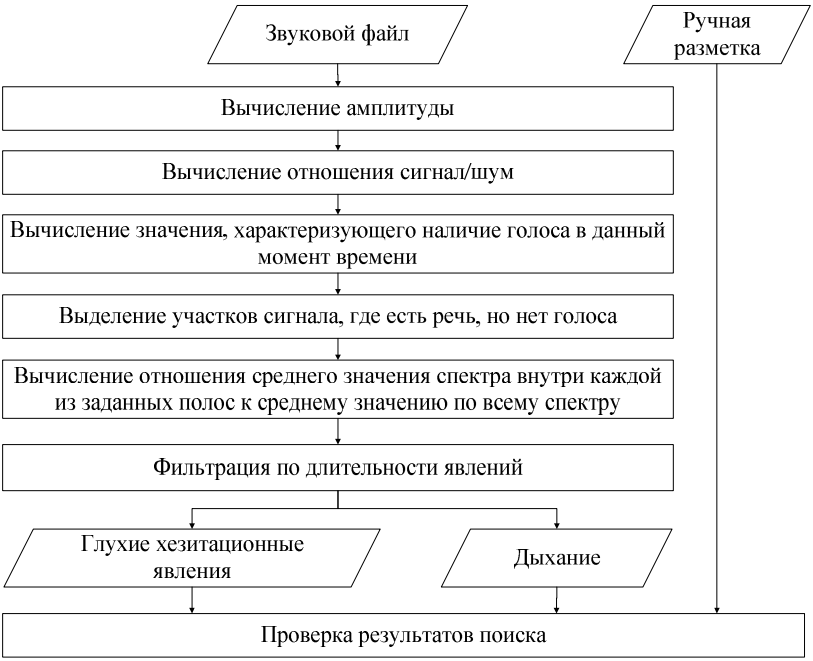

Рис.7. Схема поиска звонких хезитационных явлений.

Данные элементы представляют собой шум разной громкости и распределения мощности по спектру. Поэтому для таких явлений амплитуда сигнала больше, чем на участках с “тишиной”, и на спектро-

грамме нет участков с постоянными значениями формант. А глухие щелевые согласные имеют характерные полосы в спектре. Например, свистящие согласные имеют максимальную интенсивность спектра на частоте 4500-7000Гц.

Для поиска дыхания и удлиненных глухих щелевых согласных сначала определялся уровень шума в сигнале. Для этого производилось сглаживание значения амплитуды с использованием большой длины окна, минимум этого сглаженного ряда соответствует характерному значению для фонового шума (предполагается, что фоновый шум присутствует на всей записи и не меняется). Затем выделялись интервалы, в которых значение *SNR* не превышает 1,8 (значение подобрано экспериментально), а значение амплитуды превышает порог, равный удвоенному значению, вычисленному на предыдущем этапе. Для последовательного выделения шипящих среди найденных интервалов использовались значения отношений *R*, с порогом равным 2. Детектор артефактов и удлиненных глухих щелевых согласных выявляет шумовые явления и может выделять среди них глухие щелевые согласные (например, /s/, /ʃ/ и т.д.) путем сравнения средней амплитуды внутри определённых частотных полос со средней амплитудой по спектру, поскольку положение полосы и ее ширина постоянны. Это позволяет среди участков с шумом выделять согласные, разделяя одновременно речевые и неречевые явления.

На основе разработанных алгоритмов были созданы программные средства на языке Python. Экспериментальная проверка программных средств проводилась на основе собранного корпуса спонтанной русской речи. Точность определения звонких хезитационных явлений составила 80%. Однако анализатор находил также хезитационные удлинения, неучтенные в разметке, и сопутствующие ларингализованным участкам отрезки хезитаций и хезитационных удлинений, найденные кросскорреляционным анализом. Длительность этих отрезков была недостаточной для преодоления порога правильно найденных явлений. Так, если решить проблему с ларингализованными звуками, то процент правильного определения вырастет до 88% [34]. Точность определения артефактов составила 66%, если пренебречь тихим дыханием, которое составило больше половины ненайденных явлений, и причмокивания, нахождение которых не предполагалось в связи с ограничением на длительность элементов (минимум 200 мс).

5. Заключение. Разработанные алгоритмические и программные средства по анализу речевых сбоев в звуковом сигнале показали следующую точность определения хезитаций и дыхания: 80% для звон-

ких явлений и 66% для глухих явлений и дыхания. Необходимо отметить, что среди ненайденных явлений велика доля ларингализованных, где не срабатывает метод кросскорреляции. Основной проблемой для поиска являются ложно найденные явления, вызванные некорректным разделением гласных и сонантов между собой: неразделение или некорректное определение положения границы. Эта трудность, предположительно, вызвана большой амплитудой спектрального максимума, соответствующего ЧОТ. Изменения в формантах теряются на фоне шума в сравнении с этим неподвижным пиком. В дальнейшем исследовании предполагается поиск других критериев разделения отдельных звуков и уменьшение влияния основного максимума в спектре на значение кросскорреляционной функции.

Литература

1. *Верходанова В.О., Карпов А.А.* Моделирование речевых сбоев в системах автоматического распознавания речи // Вестн. Том. гос. ун-та. 2012. № 363. С. 10–15.
2. *Земская Е.А.* Русская разговорная речь: лингвистический анализ и проблемы обучения. М.: Русский язык, 1987.
3. *Кипяткова И.С., Верходанова В.О., Ронжин А.Л.* Сегментация паралингвистических фонационных явлений в спонтанной русской речи // Вестник Пермского университета. Российская и зарубежная филология, Вып. 2 (18), 2012. С. 17-23.
4. *Кривнова О. Ф., Чардин И. С.* Паузирование в естественной и синтезированной речи // Теория и практика речевых исследований (АРСО-99). Материалы конференции — М., 1999.
5. *Лауринавичюте А.К., Федорова О.В.* Влияние паузы хезитации на понимание синтаксической структуры предложения носителями русского языка // Материалы международной конференции «Диалог 2010». Бекасово, 2010. С. 279–284.
6. *Подлесская В.И., Кибрик А.А.* Самоисправления говорящего и другие типы речевых сбоев как объект аннотирования в корпусах устной речи // Научно-техническая информация. Сер. 2. 2007. № 2. С. 2–23.
7. *Степанова С. Б.* Общее и индивидуальное в хезитациях (на материале русской спонтанной речи) // Материалы XXXV международной филологической конференции. Фонетика. Санкт-Петербург, 2006. С. 24-32.
8. *Adams M.R.* Fluency, nonfluency, and stuttering in children // Journal of Fluency Disorders, Vol. 7, 1982. P. 171-185.
9. *Anderson A., Bader M., Bard E., Boyle E., Doherty G. M., Garrod S., Isard S., Kowtko J., McAllister J., Miller J., Sotillo C., Thompson H. S. and Weinert R.* The HCRC Map Task Corpus // Language and Speech, 34, 1991. P. 351-366.
10. *Blankenship J., Kay C.* Hesitation phenomena in English speech: A study in distribution. // Word, Vol. 20, 1964. P. 360-372.

11. *Clark H.H., Fox Tree J.E.* Using uh and um in spontaneous speaking // Cognition, Vol. 84 (1), 2002. P. 73–111.
12. *Corley M., Stewart O. W.* Hesitation disfluencies in spontaneous speech: The meaning of um // Language and Linguistics Compass, Vol. 4, 2008. P. 589-602.
13. *Fox B. A., Hayashi M., Jasperson R.* Resources and repair: a cross-linguistic study of syntax and repair // Interaction and grammar. Cambridge University Press, 1996. P.185-237.
14. *Giannini A.* Hesitation Phenomena In Spontaneous Italian // In Proceedings of the 15-th International Congress of Phonetic Sciences, Barcelona, 2003. P. 2653-2656.
15. *Howell P., Davis S., Bartrip J.* The UCLASS archive of stuttered speech. // Journal of Speech, Language, and Hearing Research, Vol. 52, 2009. P. 556–569.
16. *Kaushik M., Trinkle M., Hashemi-Sakhtsari A.* Automatic Detection and Removal of Disfluencies from Spontaneous Speech // In Proceedings of the 13-th Australasian International Conference on Speech Science and Technology (SST). Melbourne, Australia, 2010. P. 98–101.
17. *Karpov A., Markov K., Kipyatkova I., Vazhenina D., Ronzhin A.* Large vocabulary Russian speech recognition using syntactico-statistical language modeling // Speech Communication, Vol. 56, 2013. P. 213–228.
18. *Kohler K.J.* Labelled data bank of spoken standard German: the Kiel corpus of read/spontaneous speech // In Proceedings of 4-th International Conference on Spoken Language (ICSLP 96), Vol.3, 1996. P. 1938-1941.
19. *Lease M., JohnsonM., Charniak E.* Recognizing disfluencies in conversational speech // IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14(5), 2006. P. 1566–1573.
20. *Levelt W.J.M.* Monitoring and self-repair in speech // Cognition, Vol. 14, 1983. P. 41-104.
21. *Levelt W.J.M.* Speaking: From Intention to Articulation // The MIT Press, 1993. 584 p.
22. *Liu Y.* Structural Event Detection for Rich Transcription of Speech // PhD thesis. Berkeley: Purdue University and ICSI, 2004. 253 p.
23. *Liu Y., Shriberg E., Stolcke A. et al.* Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies // IEEE Transactions on Audio, Speech and Language Processing, № 14(5), 2006. P. 1526–1540.
24. *Masataka G., Katunobu I., Satoru H.* A real-time filled pause detection system for spontaneous speech Recognition // In Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech '99). Budapest, Hungary, 1999. P. 227–230.
25. *Nakatani C.H., Hirschberg J.* A corpus-based study of repair cues in spontaneous speech // Journal of the Acoustical Society of America, Vol. 95 (3), 1994. P. 1603-1616.
26. *O'Shaughnessy D.* Locating disfluencies in spontaneous speech: an acoustical analysis // In Proceedings of European Conference On Acoustics and Speech Communication and Technology. S.1., 1992. P. 2187-2190.

27. *Pätzold M., Simpson A. P.* An acoustic analysis of hesitation particles in German // In Proceedings of 13-th International Congress of Phonetic Sciences, Vol. 3, 1995. P. 512-515.
28. *Pálfy J., Pospíchal J.* Algorithms for Dysfluency Detection in Symbolic Sequences using Suffix Arrays // Text, Speech, and Dialogue. Lecture Notes in Computer Science, Vol. 8082, 2013. P. 76-83.
29. Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop // Gothenburg Papers in Theoretical Linguistics 90, Sweden, Göteborg University, 2003. P. 3–4.
30. *Shriberg E.E.* Preliminaries to a Theory of Speech Disfluencies // PhD thesis, University of California at Berkeley, 1994. 225 p.
31. *Skrelin P., Volskaya N., Kocharov D. et al.* A Fully Annotated Corpus of Russian Speech // In Proceedings of the 7-th Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 2010. P. 109–112.
32. *Snover M., Dorr B., Schwartz R.* A lexically-driven algorithm for disfluency detection // In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-Short '04), Boston, Massachusetts, USA, 2004. P. 157–160.
33. SWITCHBOARD: A User's Manual. URL: http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html (дата обращения: 05.09.2013).
34. *Verkhodanova V., Shapranov V.* Automatic Detection of Speech Disfluencies in the Spontaneous Russian Speech // Springer International Publishing Switzerland. M. Zelezny et al. (Eds.): SPECOM 2013, LNAI 8113, 2013. P. 70-77.

Верходанова Василиса Олеговна — младший научный сотрудник лаборатории речевых и многомодальных интерфейсов СПИИРАН. Область научных интересов: речевые сбои, автоматическая обработка речи, автоматическое распознавание речи. Число научных публикаций — 7. verkhodanova@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Verkhodanova Vasilisa Olegovna — junior researcher, Laboratory of Speech and Multimodal Interfaces, SPIIRAS. Research interests: speech disfluencies, automatic speech processing, automatic speech recognition. The number of publications — 7. verkhodanova@iias.spb.su; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Поддержка исследований. Данное исследование поддержано фондом РФФИ (проект № 12-06-31203-мол_а) и Комитетом по науке и высшей школе Правительства Санкт-Петербурга

Рекомендовано лабораторией речевых и многомодальных интерфейсов, заведующий лабораторией Ронжин А.Л, д.т.н., доц.
Статья поступила в редакцию 10.10.2013.

РЕФЕРАТ

Верходанова В.О. **Алгоритмы и программные средства автоматического определения речевых сбоев в звуковом сигнале.**

На сложность автоматической обработки спонтанной речи влияет ряд факторов, таких как вариативность речи или присутствие речевых сбоев различной природы. Речевые сбои могут появляться в результате воздействия внешних обстоятельств или в результате сбоя в планировании. Самыми распространенными речевыми сбоями являются паузы хезитации. Появление этих пауз говорит о том, что человеку нужно дополнительное время на формулирование следующего фрагмента высказывания. Они представляют собой семантические лакуны, часто заполненные некоторыми звуками. Другим препятствием для автоматической обработки спонтанной речи является присутствие артефактов – физиологических явлений как смех, кашель, вздох.

В статье рассматриваются разные виды речевых сбоев, причины их появления, а также описывается алгоритм и программные средства их автоматического определения. Для исследования этих явлений был собран корпус спонтанных диалогов двух типов: описания маршрута по карте (т.н. map-task) и нахождения свободного времени по расписанию (т.н. appointment-task). Корпус был вручную аннотирован в программе Wave Assistant. В ходе аннотации были размечены такие элементы, как заполненные паузы хезитации (например, [ə], [ɐ]), артефакты речи (например, смех, вздох), самокоррекции и фальстарты, а также слова и словосочетания, заполняющие паузы. В данном исследовании использовалась часть корпуса длиной 50 минут - 18 диалогов длительностью от 1,5 до 5 минут. Все записи были сделаны в Санкт-Петербурге в конце 2012 – начале 2013 года.

Был разработан алгоритм автоматического определения речевых сбоев и артефактов в звуковом сигнале. Для выделения звонких хезитационных явления использовался кросскорреляционный метод, а для выделения глухих хезитационных явлений – метод полосовой фильтрации.

Предложенный алгоритм автоматического определения речевых сбоев, опирающийся на акустические параметры сигнала, был реализован в виде программы по выявлению хезитационных явлений и дыхания. Проведенные эксперименты показали, что звонкие хезитационные явления выделяются с точностью 80%, глухие хезитационные явления и дыхание - с точностью 66%. В дальнейшем планируется исследовать проблемы поиска ларингализованных звуков и точного определения границ речевых сбоев в звуковом сигнале.

SUMMARY

Verkhodanova V.O. **Algorithms and Software for Automatic Detection of Speech Disfluencies in an Audio Signal**

A number of factors such as speech variation and different kinds of speech disfluences has a bad influence on automatic speech processing. Speech disfluencies may occur due to influence of external situation or due to disturbances in the discourse planning. The most common speech disfluencies are filled pauses (hesitational pauses). Their appearance indicates that the speaker needs an additional time to formulate the next piece of utterance. They are semantic lacunas that are often filled with certain sounds. Another obstacle for automatic spontaneous speech processing is the presence of speech artefacts such as laugh, cough or sigh.

In this article different types of speech disfluencies and their causes are presented, as well as the algorithm and the software for their automatic detection based on the acoustical parameters. For the purpose of study a corpus of spontaneous map-task and appointment-task Russian dialogs was collected. This corpus was manually annotated by means of Wave Assistant. During annotation such phenomena as filled pauses ([ə], [ɐ]), artefacts (laugh, sigh), self-repairs and false-starts were marked. In this study a part of this corpus was used: 18 dialogs from 1.5 to 5 minutes long. All the recordings were made in St.Petersburg in the end of 2012 – beginning of 2013.

The algorithm for automatic detection of speech disfluencies and artefacts in an audio signal was developed. To detect voiced hesitational phenomena a cross-correlation method was used, and to detect unvoiced hesitational phenomena – a method of band-filtering. The proposed algorithm was implemented as a Python program for detecting hesitational phenomena and breath. The experiment showed that the detection accuracy of voiced hesitations and lengthenings is 80% and of unvoiced ones and breath it was 66%. Further it is planned to solve the occurred problems of laryngealized sounds and of inaccurate division between sounds.