С./Pp. 104–118 ВОПРОСЫ ЯЗЫКОЗНАНИЯ 2018. № 6
Voprosy Jazykoznanija
DOI: 10.31857/S0373658X0002022-3

Автоматическое определение вокализованных хезитаций в русской речи

Василиса Олеговна Верходанова[@], Владимир Владимирович Шапранов, Ирина Сергеевна Кипяткова, Алексей Анатольевич Карпов

Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, Россия; [@] vass.verkhodanova@gmail.com

Аннотация: Статья посвящена проблеме автоматического определения наиболее частотных речевых сбоев в русской речи — хезитаций. Описываются акустические свойства русских хезитационных явлений, а также приводится анализ разных методов их автоматического определения. Акустический анализ показал, что вокализации в русской речи имеют тенденцию к централизации, а также по-разному влияют на окружающий их контекст в зависимости от речевого жанра. Проведенные эксперименты по программному определению хезитационных явлений в русской речи показали эффективность и адекватность подходов, опирающихся только на акустическую информацию. В результате лучшим оказался метод опорных векторов, при котором взвешенное гармоническое среднее точности и полноты определения хезитационных явлений достигает 56 %.

Ключевые слова: автоматическая обработка речи, машинное обучение, паралингвистический анализ речи, русская речь, хезитации

Для цитирования: Верходанова О. В., Шапранов В. В., Кипяткова И. С., Карпов А. А. Автоматическое определение вокализованных хезитаций в русской речи // Вопросы языкознания. 2018. № 6. С. 104–118. DOI: 10.31857/S0373658X0002022-3.

Благодарности: Данное исследование проводится при поддержке РФФИ (проекты №№ 15-06-04465 и 18-07-01407), Совета по грантам Президента РФ (проекты №№ МК-1000.2017.8 и МД-254.2017.8), а также бюджетной темы № 0073-2018-0002.

Automatic detection of vocalized hesitations in Russian speech

Vasilisa O. Verkhodanova[@], Vladimir V. Shapranov, Irina S. Kipyatkova, Alexey A. Karpov

St. Petersburg Institute for Informatics and Automation, Russian Academy of Sciences, St. Petersburg, Russia; [@] vass.verkhodanova@gmail.com

Abstract: The article is focused on the automatic detection of the most frequent speech disfluencies in Russian speech — hesitations. Authors describe the acoustic features of Russian hesitations as well as analyze the different methods of hesitation detection. Results of acoustic analysis have shown that hesitations in Russian speech tend to be centralized, and dependent on the speech genre influence the context differently. Experiments on computerized detection of hesitations in Russian speech confirmed the efficiency and adequacy of the approaches based on acoustic information alone. Support vector machines method yielded the best results with the weighted harmonic mean of precision and recall reaching 56 %.

Keywords: automatic speech processing, hesitations, machine learning, paralinguistic speech analysis, Russian speech

For citation: Verkhodanova V. O., Shapranov V. V., Kipyatkova I. S., Karpov A. A. Automatic detection of vocalized hesitations in Russian speech. *Voprosy Jazykoznanija.* 2018. No. 6. Pp. 104–118. DOI: 10.31857/S0373658X0002022-3.

Acknowledgements: The research is supported by RFBR (projects No. 15-06-04465 and 18-07-01407), the Council for grants of the President of the Russian Federation (projects No. МК-1000.2017.8 and МД-254.2017.8), and the budget theme No. 0073-2018-0002.

Введение

В последние годы речевые технологии развиваются чрезвычайно быстро: в общественных местах используются справочные диалоговые системы, появляется программное обеспечение для помощи слепым и слабовидящим, разрабатываются голосовые помощники, позволяющие управлять приложениями в телефоне (например, голосовые помощники «Алиса» от Яндекса, Siri от Apple, Google Now, Cortana от Microsoft, Alexa от Amazon, Facebook M и др.). Чаще всего речевые технологии разрабатываются на основе подготовленной и читаемой речи. Однако практически вся та речь, которую мы ежедневно порождаем и воспринимаем, является спонтанной, и процесс автоматической обработки такого материала приводит к ряду сложностей.

В спонтанной речи говорящему приходится одновременно решать несколько сложных когнитивных задач за короткое время. Человеку нужно сформировать высказывание, выбрать правильную лингвистическую форму: подобрать нужные слова, выражения, грамматические формы и т. д. Поэтому процесс порождения спонтанной речи сопровождается различными ошибками, которые часто называют речевыми сбоями. Несмотря на актуальность вопроса и разносторонние исследования этих явлений, общепринятая терминология в этой области до сих пор не сложилась. В англоязычной литературе можно встретить термины «speech disfluencies», «non-fluency», «turn-holding devices» и др., в русскоязычных работах помимо «речевых сбоев» можно встретить такие термины как «внеязыковые элементы речи» или «паралингвистические явления» [DiSS'03 2003]. Речевые сбои имеют разную природу возникновения. К ним относятся, например, самоисправления или самокоррекции (*Я приду… **приеду** завтра в пять*), фальстарты (*Здра… **привет!***), повторения (*Я хотел бы... **хотел бы** чашечку кофе*), хезитационные явления (ХЯ), включающие в себя заполненные паузы (*У меня поезд завтра **э-э** с Курского вокзала*) и удлинения звуков (*Вы не могли б**ы-ы** повторить?*), и ряд других нарушений, появляющихся в связной и гладкой речи. Классификация речевых сбоев тоже не всегда однозначна. Так, ХЯ рассматривались разными исследователями по-разному [Shriberg 1994]: некоторые авторы классифицировали их вместе с артефактами (смехом, кашлем), другие же считали их языковыми элементами, объединяя с сочинительными союзами или с дискурсивными маркерами.

Причина возникновения речевых сбоев тоже не всегда однозначна: исследователи рассматривают их и как проблему поиска слов [Eisler 1968], и как сложности концептуализации на значимых границах дискурса [Chafe 1980]. В спонтанной речи сбои появляются часто: вплоть до трети высказываний содержит хотя бы один речевой сбой [Shriberg 1994]. Например, в американском английском на каждые 100 слов в среднем приходится шесть со сбоями [Shriberg 1994].

В русской речи заполненные паузы появляются примерно один раз на четыре слова, при этом частота появления сбоев одинакова как внутри синтагм, так и на границах дискурса [Кибрик, Подлесская 2014]. Несмотря на то, что данные о хезитациях различаются в зависимости от языка, жанра речи и социальной группы говорящего, в среднем на каждые 100 слогов приходится несколько заполненных пауз; они также являются самыми частотными речевыми сбоями [O'Connell, Kowal 2004]. Согласно [Stolcke et al. 1998], заполненные паузы

составляют около 40 % всех речевых сбоев в корпусе телефонных диалогов Switchboard [Godfrey et al. 1992]. В португальском корпусе лекций LECTRA 1,8 % всех слов — это заполненные паузы, при этом они составляют 30 % всех сбоев [Medeiros et al. 2013a].

Хезитационным явлениям свойственны как универсальные характеристики, так и специфичные для языков и жанров. Обычно заполненные хезитации представляют собой вокализации, в редких случаях — удлинения согласных, как, например, в армянском [Хуршудян 2005]. Хезитации обычно произносятся с минимальными усилиями, что соответствует принципу артикуляторной экономии [Stepanova 2007]. Однако некоторые исследования показали, что фонологическая система языка может влиять на качество хезитационных вокализаций [Giannini 2003]. Даже такие универсальные свойства ХЯ, как появление ларингализации в удлинениях, проявляется по-разному в различных языках. Например, было показано, что в финской речи ларингализация может обозначать смену говорящего в диалоге [Ogden 2001], в то время как в английской речи этого не происходит [Shriberg 2001].

ХЯ вместе с другими речевыми сбоями ранее часто рассматривались как речевой мусор и ошибки, однако они являются значимой составляющей естественного диалога [Кибрик, Подлесская 2014; Shriberg 2005]. Они выполняют различные функции: могут работать маркерами поиска, показывая, что человек еще не закончил формулировать высказывание и ему требуется дополнительное время на подбор слова [Shriberg 2005; Ogden 2001]. Также они могут быть дейктическими или дискурсивными маркерами (например, сигнализируя о начале или конце высказывания), или даже служить ритмообразующими элементами (подробнее см. [Богданова-Бегларян 2014; 2016]).

В области патологий речи анализ речевых сбоев не менее важен. Так, в [Arbisi-Kelm, Jun 2005] было проведено сравнение просодических свойств речевых сбоев в речи людей с заиканием и без заикания. Анализ спонтанных монологов восьми человек (четырех — с заиканием и четырех — без него) показал, что в речи заикающихся людей речевые сбои встречаются чаще, а появление сбоев по-разному влияет на мелодический рисунок контекста и фразы в группах с заиканием и без. В [Esposito et al. 2016] было проведено сравнение ХЯ в речи здоровых людей и людей с депрессией. Авторы показали, что в этих двух группах качественные и количественные характеристики ХЯ различаются. Таким образом, распознавание ХЯ может не только улучшить работу систем автоматического распознавания речи (АРР), но и способствовать определению речевых и даже психических проблем, а также степени владения языком.

Данная статья посвящена проблеме автоматического определения ХЯ в русской речи, основанного на акустическом анализе. В разделе 1 мы кратко обозреваем современное состояние проблемы определения ХЯ в мире. Раздел 2 посвящен акустическому анализу ХЯ, а раздел 3 описывает результаты экспериментов по автоматическому определению хезитаций на разнообразном материале русской речи.

1. Аналитический обзор методов автоматического определения хезитаций

Автоматическая обработка спонтанной речи связана со многими нерешенными проблемами, такими как распознавание эмоций в речи [Kaya, Karpov 2018] и иных паралингвистических явлений [Kaya et al. 2017] и необходимостью учитывать речевые сбои, в частности ХЯ. К проблеме распознавания этих явлений пытались подходить с позиций разных дисциплин. В компьютерной лингвистике анализ речевых сбоев иногда включают в системы синтаксических парсеров [Ferreira et al. 2004], но гораздо чаще подобный анализ является частью систем АРР [Liu et al. 2006; Кипяткова, Карпов 2016]. ХЯ, как и речевые сбои в целом, всегда были препятствием для автоматической обработки спонтанной речи. Поскольку ХЯ могут появиться в любой момент и в любом месте высказывания, системы распознают

их неправильно и ошибаются при классификации соседних слов [Medeiros et al. 2013a; Shriberg 2005; Audhkhasi et al. 2009; Goto et al. 1999].

Недавний всплеск интереса к проблеме автоматического определения речевых сбоев и ХЯ произошел в 2013 г., когда в рамках международной конференции INTERSPEECH проводился конкурс по определению некоторых паралингвистических явлений в речи, в частности, по автоматическому определению заполняющих элементов (хезитаций, слов-паразитов и т. д.) [ComParE 2013; Schuller et al. 2013]. Победителем этого конкурса стала система, в которой для определения указанных явлений был построен классификатор на основе глубинных нейронных сетей [Gupta et al. 2013].

Чаще всего автоматическое распознавание речевых сбоев направлено на лексически выраженные элементы, такие как фальстарты, исправления и др. Эту задачу часто решают с помощью применения методов распознавания шаблонов (sequence tagging) или с помощью построения синтаксических моделей. Однако, поскольку ХЯ являются наиболее частотными речевыми сбоями, многие работы посвящены распознаванию и определению только этого типа речевых сбоев.

В одной из первых работ по автоматическому определению ХЯ [Shriberg et al. 1997] было показано, что длительности и уровня падения частоты основного тона (ЧОТ), а также расстояния до абсолютной паузы было достаточно для классификации ХЯ. Однако для уверенного автоматического распознавания ХЯ этих характеристик недостаточно [Verkhodanova et al. 2016].

Другой ранней работой по распознаванию хезитаций в речевом сигнале стала работа японских исследователей [Goto et al. 1999]. Авторы использовали два признака для определения ХЯ в сигнале: небольшое изменение частоты основного тона и слабое искажение спектральной огибающей. Метод был опробован на 100 предложениях из корпуса японской разговорной речи, каждое из которых содержало хотя бы одно ХЯ. Полученные результаты показали 91,5 % точности (precision) и 84,9 % полноты (recall) распознавания ХЯ. Однако позже авторы признали, что столь высокие значения получились во многом благодаря однородности материала, поскольку в их корпусе не было низких мужских голосов.

Последующие работы по автоматической обработке ХЯ использовали и используют более сложные и разнообразные наборы признаков, в поисках «золотого стандарта» комбинируя разнообразные акустические признаки, результаты работы систем APP или даже лексический контекст. Так, в [Stouten, Martens 2003] была предложена система определения ХЯ для улучшения работы системы APP. В качестве классификатора использовался многослойный перцептрон с одним выходом, а набор признаков включал в себя сегментную длительность, стабильные интервалы, наличие паузы до и после ХЯ. Система была протестирована на материале фламандского разговорного языка из голландского корпуса, в результате точность определения составила 85 %, а полнота — 70 %.

В [Medeiros et al. 2013a] авторы фокусировались на распознавании заполненных пауз на основе акустических и просодических, а также некоторых лексических свойств речи. Эксперименты проводились на корпусе европейского португальского языка LECTRA. Были протестированы несколько методов машинного обучения, но наилучшие результаты показал метод классифицирующих и регрессионных деревьев. Для задачи определения слов внутри участков с речевыми сбоями результаты составили около 91 % точности и 37 % полноты, при этом в качестве одного из признаков использовались сами заполненные паузы. Без учета этого признака результаты снижались до 66 % для точности и до 20 % для полноты. Дальнейшие эксперименты по определению ХЯ в европейском португальском проводились с использованием просодических и лексических признаков, полученных в результате работы системы APP. В итоге, лучшие результаты показал алгоритм классификации J48: значение F-меры составило 61 % [Medeiros et al. 2013b].

В [Prylipko et al. 2014] был представлен алгоритм определения заполненных пауз с помощью метода опорных векторов. В экспериментах также использовался фильтр Гаусса для получения контекстной временно́й информации и морфологическое открытие для

фильтрации ложно-положительных срабатываний. Использовался набор акустических признаков, схожий с предложенным в соревнованиях по компьютерной паралингвистике [ComParE 2013; Schuller et al. 2013]. Для экспериментов был выбран немецкий корпус LAST MINUTE, содержащий многомодальные записи 133 носителей немецкого языка. В результате работы системы точность определения заполненных пауз составила 70 %, полнота — 55 %.

В России исследования ХЯ в русской речи проводились в разных направлениях: от анализа ХЯ в дискурсивных исследованиях [Подлесская, Кибрик 2007; Богданова-Бегларян 2016] до изучения их акустических свойств [Stepanova 2007; Verkhodanova et al. 2017] и разработки методов их обнаружения [Verkhodanova et al. 2016; 2017]. Подробнее об акустических исследованиях ХЯ и об экспериментах по их определению говорится в разделе 2.

2. Акустический анализ ХЯ

2.1. Речевой материал

Обычно для исследования речевых сбоев используются корпуса спонтанной речи с многоуровневой аннотацией. Помимо такой информации, как фонемы, слова и синтагмы, дополнительно отмечаются речевые сбои. Такая транскрипция называется «богатой» (Rich Transcription) [Liu 2004]. Примером может служить корпус английских телефонных разговоров English CTS Treebank [English CTS], в аннотации которого, помимо прочего, учитываются речевые сбои и дискурсивные маркеры. Для русской речи можно упомянуть корпус, разработанный на кафедре фонетики СПбГУ, разметка которого проводилась на шести уровнях и учитывала различные виды фонетической и просодической информации [Skrelin et al. 2010].

Для задач исследования ХЯ в русской речи авторами данного исследования использовался аудиоматериал из четырех корпусов, содержащий записи речи различных жанров в разных коммуникативных ситуациях; кроме того, записи различаются по качеству.

1. Корпус спонтанных диалогов. Записи, вошедшие в этот корпус, были сделаны в СПИИРАН в Санкт-Петербурге в 2012–2013 гг. В сеансах записи принимали участие 12 человек: шесть девушек и шесть юношей в возрасте от 17 до 23 лет как технических, так и гуманитарных специальностей (по три представителя каждого пола) с полным или неоконченным высшим образованием. Участники выполняли два задания: 1) описать маршрут по карте и 2) найти общее время для встречи в соответствии с индивидуальным расписанием. Запись проводилась в звукоизолированной комнате с использованием мобильных устройств (планшетов) Samsung Galaxy Tab 2, бесплатным приложением Smart Voice Recorder. Среди размеченных явлений встретились 492 хезитации (222 заполненные паузы и 270 удлинений) [Верходанова 2013].

Дополнительно для ряда экспериментов были использованы записи спонтанных диалогов, не вошедших в оригинальный корпус, поскольку эта часть не была сбалансирована по полу и специализации испытуемых. Сюда вошли 39 диалогов между парами студентов (24 юноши и две девушки) с техническим образованием. Студенты выполняли те же задания описания маршрута по карте и нахождения общего времени по расписаниям. Всего в этих записях было размечено 412 хезитаций (211 заполненных пауз и 201 удлинение).

2. Мультиязычная речевая база данных из открытых источников. Монологическая спонтанная речь в материале представлена рядом записей из мультиязычной базы данных американского Бингемптонского университета [Zahorian et al. 2011]. Эта база включает в себя более 30 часов записей на трех языках — английском, китайском и русском. В нее вошли в частности 900 записей из открытых веб-сайтов, таких как youtube.com

и rutube.com, поэтому уровень качества и степени зашумленности материала варьируется. Средняя длительность аудиовидеофайлов составила семь минут. Русскоязычная часть корпуса включает 300 записей 158 говорящих, что приблизительно соответствует 35 часам речи. Эти записи делятся на два жанра: «Формальные презентации» и «Спонтанные разговоры». Подобное разделение соответствует условиям записи: первая представляет собой записи заранее подготовленной речи в более-менее тихих условиях, вторая — в основном зашумленные спонтанные монологи и интервью. Русскоязычная часть спонтанных разговоров состоит из 91 записи (около 10 часов) 53 говорящих [Zahorian et al. 2011]. Для исследований ХЯ были выбраны шесть спонтанных разговоров шести говорящих (трех женщин и троих мужчин), которые были размечены экспертным путем на речевые сбои. При аннотации были выделены 284 хезитации (198 заполненных пауз и 86 удлинений).

3. Аудиозаписи докладов семинара АРЗ. Частично подготовленная монологическая речь представлена научными докладами, сделанными в ходе семинара АРЗ [АРЗ 2011], которые были записаны в СПИИРАН в 2011 г. Шесть записанных докладов были посвящены анализу и обработке разговорной речи. Дикторы (три женщины и трое мужчин) не опирались на письменный текст, поэтому данную речь можно рассматривать как приближенную к спонтанной. Записи были вручную аннотированы экспертом с учетом речевых сбоев, изобилие которых говорит о схожести данного типа речи со спонтанной. Всего было выделено 951 ХЯ: 751 заполненная пауза и 210 удлинений.

4. Записи из приложения № 5 к Бюллетеню Фонетического фонда (БФФ). Все аудиозаписи были сделаны в Москве, кроме одной, которая выполнена в Праге, в конце 1970-х и в 1980-е гг. Записывалась речь в ходе докладов ученых на семинарах, конференциях, заседаниях. Докладчики выступали без опоры на письменный текст, иногда доклады переходили в дискуссии. Большинство информантов имеет ученую степень, и для всех русский язык является родным. Все 12 информантов (шесть мужчин и шесть женщин) являются носителями литературного русского языка. В данном корпусе содержатся записи, отражающие проблематику нескольких научных дисциплин: лингвистики, методики обучения языку, логики, психологии, науковедения. Всего было размечено 285 хезитаций (225 заполненных пауз и 60 удлинений).

Таким образом, объем всех четырех корпусов составил приблизительно 5,7 часов, а общее количество размеченных хезитаций — 2422.

2.2. Акустический анализ ХЯ в русской речи

Как было показано ранее [Verkhodanova, Shapranov 2015], большей части ХЯ в русской речи, также как и ХЯ в других языках [O'Shaughnessy 1992], свойственна стабильность спектральных характеристик и длительность более 150 мс.

В работе [Stepanova 2007] были проанализированы некоторые спектральные характеристики хезитаций (первая и вторая форманты) в спонтанных монологах 10 носителей русского языка. Сравнение характеристик хезитаций и реализаций ударных гласных фонем /а/ и /э/ показало, что хезитации подвержены централизации в рамках пространства гласных при значительной вариативности произнесений для различных дикторов. Это также соответствует одному из основных типов редукции гласных вообще [Barnes 2006].

На материале всех четырех корпусов был проведен подробный акустический анализ встреченных хезитационных явлений, который подтвердил тенденцию к централизации гласных в вокализованных хезитациях (рис. 1). Хезитации классифицировались экспертами по слуховому впечатлению; ниже на диаграммах показано сравнение распределений самых частотных хезитаций с соответствующими заполняющим звукам фонемами.

Анализ распределений двух самых частотных вокализованных ХЯ со схожими ударными фонемами /ɜ́/ и /а́/ показал, что, несмотря на некоторые индивидуальные особенности

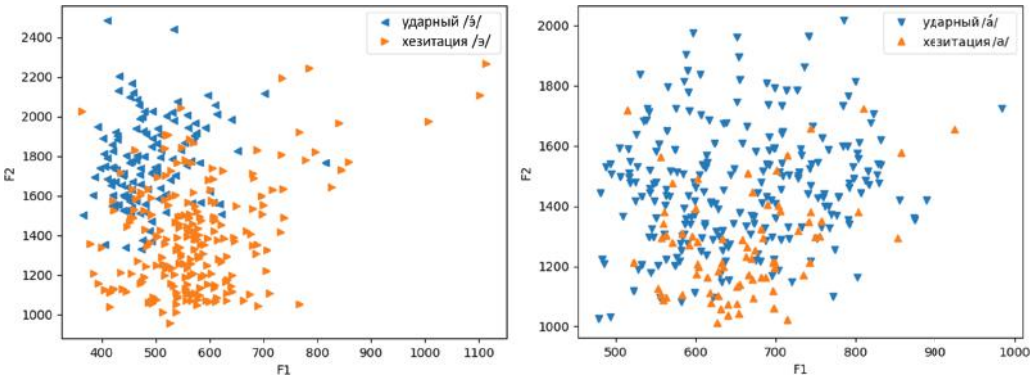

Рис. 1. Сравнение распределений /э/-образных ХЯ с ударными фонемами /э́/ (слева) и сравнение /а/-образных ХЯ с ударными фонемами /а́/ (справа).
По оси абсцисс — первая форманта (F1), по оси ординат — вторая форманта (F2).

дикторов, общая тенденция централизации гласных очевидна как для первой, так и для второй форманты (F1 и F2) в /э/-образных ХЯ. Эта же тенденция прослеживается для ХЯ, близких к /а/, хотя разброс значений для формант в этом случае значительно больше. С одной стороны, это свидетельствует о возможном неоднородном восприятии и разметке экспертами хезитаций, заполненных фонемой /а/, а с другой стороны, сигнализирует о более неоднородной спектральной картине этих явлений и, возможно, об особенностях использования хезитаций, заполненных фонемой /а/.

На материале всех четырех подкорпусов был проведен акустический анализ встреченных хезитационных явлений. Распределение длительности хезитаций в объединенном корпусе представлено на рис. 2. Средняя длительность хезитации составила 380 мс, причем значения длительности отдельных экземпляров варьировались от 6 мс до 2,3 с.

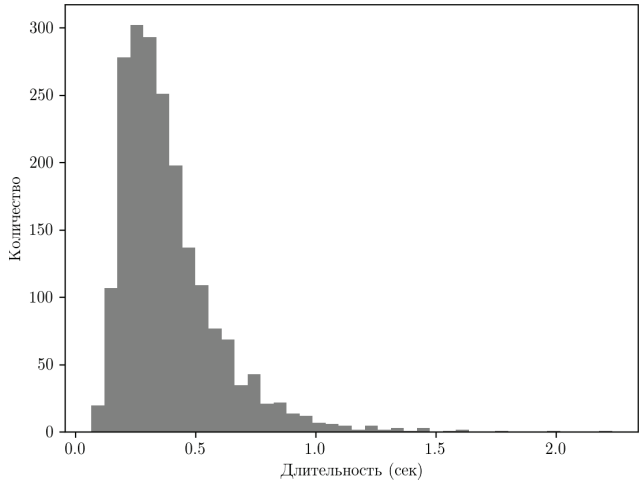

Рис. 2. Распределение длительностей хезитаций во всем корпусе

Самые частотные явления представлены на рис. 3. Наиболее частотным явлением, как и предполагалось, стали заполненные паузы. Их доля составила 43 % от всех ХЯ в корпусе. Самой распространенной заполненной паузой стал долгий /э/-образный звук, обозначавшийся в корпусе как «h.e» (хезитация типа /э/), самым распространенным удлинением стало

удлинение фонемы /и/, которое встречалось, в основном, в предлогах или лексических заполнителях пауз (например, «и... вот»).

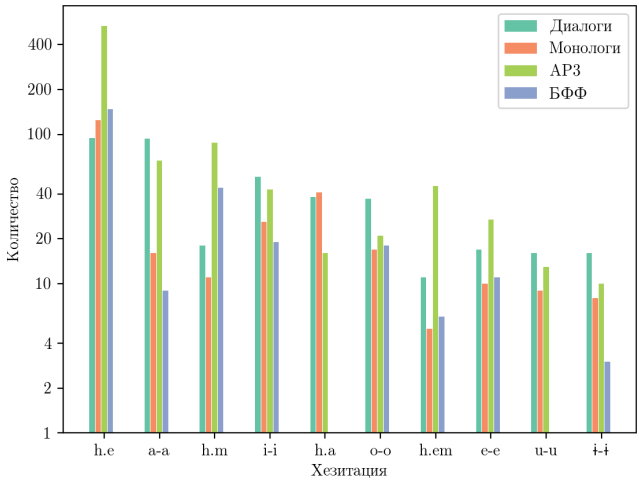

Рис. 3. Распределение десяти самых частотных ХЯ в различных частях объединенного корпуса. Условные обозначения: h.e — /э/-образная хезитация, h.a — /а/-образная хезитация, h.m — /м/-образная хезитация, h.em — хезитация, переходящая из звука [э] в звук [м] («эм»); a-a, i-i, o-o, e-e, u-u, ɨ-ɨ — удлинения фонем /а/, /и/, /о/, /э/, /у/ и /ы/[1] соответственно.

Как уже было отмечено ранее, стабильность ЧОТ и ее постепенное падение является универсальным свойством ХЯ и мало зависит от языка. С другой стороны, мало известно о том, как ХЯ влияют на окружающую речь с точки зрения акустики, и еще меньше известно о поведении ХЯ в разных речевых жанрах. Подобное исследование было проведено для европейского варианта португальского языка [Moniz et al. 2014], где авторы показали, что распределение ХЯ для различных речевых жанров различается. Также было проверено, насколько могут различаться в разных жанрах просодические маркеры контраста, например, темп речи, средняя длительность предложений со сбоями и без сбоев и др. Так, диалоги оказались просодически более разнообразными, чем лекции, в которых было в среднем меньше слов в единицу времени.

Анализ акустико-просодической вариативности контекста вокруг ХЯ в различных жанрах русской речи был впоследствии проведен и на русскоязычном материале [Verkhodanova et al. 2017]. Сравнивались значения ЧОТ (F0) и уровня энергии для монологов и диалогов на материале объединенных корпусов, описанных выше. Эти два параметра были выбраны как содержащие важную просодическую информацию, которая доступна напрямую из речевого сигнала. Было проведено сравнение значений ЧОТ и энергии для следующих пар: хезитация и левый контекст, хезитация и правый контекст в монологической и диалогической речи. Контекст для диалогической части общего корпуса брался из экспертной разметки, в то время как для монологической части он рассчитывался автоматически путем проверки наличия ЧОТ вокруг хезитации в окне 400 мс.

Данное сравнение показало статистически значимые различия между отношением распределения значений ЧОТ в хезитациях и в их контекстах (как для левого, так и для правого) в монологах и диалогах (см. рис. 4 и 5). Что касается энергии, то значимые различия были найдены только для отношения распределения ХЯ и правого контекста в монологах и диалогах (см. рис. 6).

[1] В данном исследовании /ы/ рассматривалась как самостоятельная фонема.

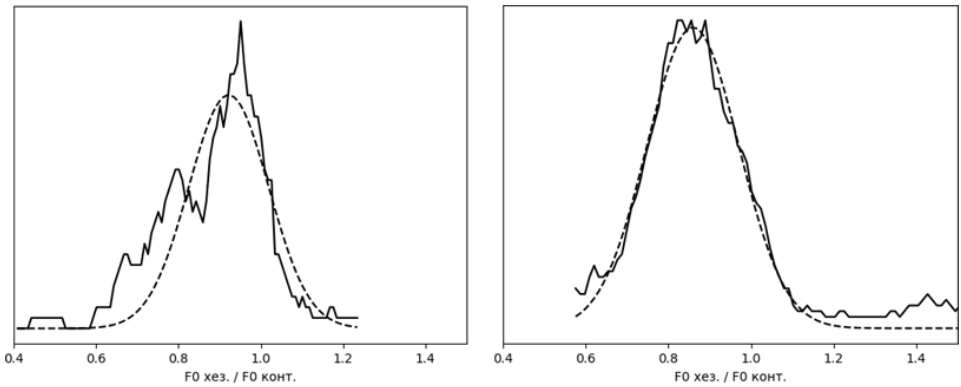

Рис. 4. Различия в ЧОТ между хезитацией и правым контекстом для диалогов (слева) и монологов (справа) в русской речи[2]

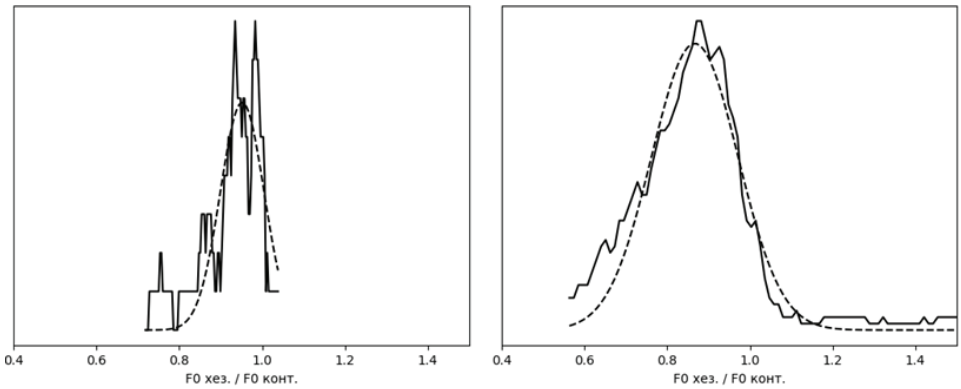

Рис. 5. Различия в распределении ЧОТ между хезитацией и левым контекстом для диалогов (слева) и монологов (справа)

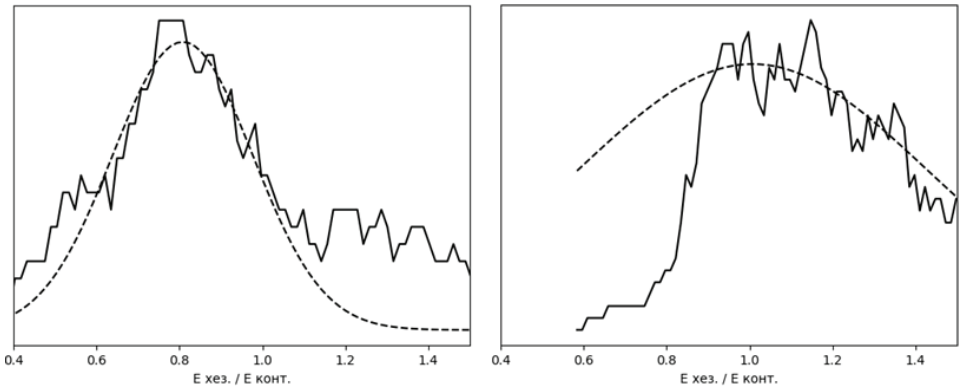

Рис. 6. Различия в уровне энергии между хезитацией и правым контекстом для диалогов (слева) и монологов (справа)

[2] На рис. 4–6 пунктирная линия показывает аппроксимацию нормальным распределением.

Эти данные согласуются с результатами по европейскому португальскому, [Moniz et al. 2014], где наибольший скачок ЧОТ в месте перехода от речевого сбоя к обычной речи был характерен именно для ХЯ. В русской монологической речи просодические маркеры чаще встречаются в контекстах хезитаций, чем в диалогической речи. На рис. 3 и рис. 4 видно, что для диалогической речи различия в отношении ЧОТ между хезитациями и контекстами менее очевидны, чем в монологической речи. Однако распределение значений энергии в европейском португальском выбивается из общей тенденции более выраженного просодического маркирования речевых сбоев в монологической речи. Для русской речи распределение значений энергии вписывается в общую картину. Возможно, такие различия в данных обусловлены не столько разницей в стратегиях маркирования участков со сбоями в различных языках, сколько методом измерения энергии. В статье [Moniz et al. 2014] авторы измеряли среднее значение энергии внутри контекста и хезитации, а для данного исследования использовалась аппроксимация распределения энергии нормальным распределением для уменьшения влияния внешних шумов.

3. Эксперименты по определению ХЯ в русской речи

3.1. Основанные на правилах подходы к задаче определения ХЯ в русской речи

Пилотный эксперимент, проверяющий возможность простого определения ХЯ в русской речи, описан в работе [Verkhodanova et al. 2016]. Эксперименты проводились на корпусе спонтанных диалогов. Алгоритм учитывал значения длительности, первых трех формант, энергии и спектральную стабильность во всем корпусе. Релевантность использования этих акустических свойств для определения ХЯ в других языках была показана в работах [Audhkhasi et al. 2009; Garg, Ward 2006]. В качестве параметров в экспериментах использовались стандартные отклонения ЧОТ (F0), первой форманты (F1) и энергии, поскольку именно эти значения оказались наиболее стабильными внутри хезитации. F-мера была выбрана как мера оценки результатов, поскольку она позволяет объединить точность и полноту в одной усредненной величине. Эта мера определяется как взвешенное гармоническое среднее точности и полноты. Для максимизации значения F-меры подбирались оптимальные значения параметров a и b для следующего критерия, где X — стандартное отклонение логарифма ЧОТ, а Y — стандартное отклонение логарифма энергии:

$$C = aX + bY < 1$$

Оптимальные значения параметров выбирались для максимизации F-меры в рамках задачи нахождения фреймов, относящихся к ХЯ. Дополнительным порогом служило значение стандартного отклонения логарифма первой форманты. В результате было достигнуто значение F-меры 41 %.

На следующем этапе экспериментов [Verkhodanova, Shapranov 2015] критерий был обобщен для большего числа параметров, а также был изменен процесс максимизации параметров:

$$C = \sum_{n} w_n V_n < 1$$

В формуле w_n — веса для значений всех параметров (V_n); в качестве параметров взяты стандартные отклонения энергии и первых трех формант ($log(E)$, $log(F_N)$); а минимальное

значение уровня энергии использовалось в качестве порога. Кроме этого, к корпусу спонтанных диалогов был добавлен материал из мультиязычной речевой базы данных из открытых источников. Максимизация значения F-меры производилась методом градиентного спуска [Snyman 2005], в результате было получено значение F-меры 46 %.

Для этих двух подходов ошибки работы метода были связаны с невозможностью обработать ларингализованные ХЯ, с неэффективной обработкой нестабильных контуров ЧОТ, а также со случаями наложения голосов и высокого уровня шума. Таким образом, несмотря на то, что подобные классификаторы учитывают значительную часть ХЯ, возможность эффективно учитывать остальные ХЯ с их помощью вызывает сомнения.

3.2. Подходы к определению ХЯ в русской речи, основанные на данных

Применение алгоритмов машинного обучения к задаче определения ХЯ в русской речи началось с [Verkhodanova et al. 2016], где эксперименты по определению ХЯ проводились с использованием нейросети на основе метода экстремального обучения (ELM — Extreme Learning Machines). Этот подход представляет собой искусственную нейронную сеть, ориентированную на решение задач классификации и регрессии. В экспериментах использовалась реализация ELM на языке Python, описанная в [Akusok et al. 2015]. Сеть состояла из 600 сигмоидальных нейронов, и в качестве признаков был взят набор из 20 стандартных отклонений (первые три форманты, энергия, вероятность вокализации и ее производная и 14 мел-частотных кепстральных коэффициентов MFCC), а также трех средних значений (энергии, вероятности вокализации и ее производной). Значения формант были получены с помощью открытого программного обеспечения Praat [Boersma, Weenink 2016], остальные параметры — с помощью программной библиотеки openSMILE [Eyben et al. 2010]. Используемый материал (спонтанные диалоги, записи из мультиязычной речевой базы данных, а также приложение к БФФ № 5) был разделен на тестовый и обучающий подкорпуса. Случайно выбранные 10 % были отведены на тестовый корпус, остальные — на обучающий. Процедура кросс-валидации проводилась 10 раз (10-fold cross validation), порождая при этом различные наборы тестовых и обучающих данных. Поскольку объемы ХЯ и основной речи не были сбалансированы внутри корпуса, обучающий подкорпус прошел процедуру уменьшения объема данных (downsampling), чтобы избежать смещения оценки в сторону класса «речь» [Prylipko et al. 2014]. После этого происходило обучение классификатора. Для каждого фрагмента метод ELM выдает число, которое при превышении определенного порога характеризует этот фрагмент как принадлежащий к ХЯ. Пороговое значение выбиралось методом поиска по сетке, оптимизирующим значение F-меры на обучающем корпусе. В результате было получено значение F-меры, равное 42 %.

В последних экспериментах по применению машинного обучения для определения ХЯ в русской речи использовался метод опорных векторов (SVM — Support Vector Machines) [Verkhodanova, Shapranov 2016a; Verkhodanova et al. 2016, 2017]. Актуальность и эффективность этого метода была показана в [Prylipko et al. 2014]. По сравнению с ELM, метод опорных векторов позволяет получить лучшее значение точности распознавания при лучшем значении F-меры. Для экспериментов использовалась реализация метода с полиномиальным ядром, позволяющая оценивать вероятность наличия ХЯ с помощью классификации методом опорных векторов (C-Support Vector Classification). Данная реализация входит в программную библиотеку [Scikit-Learn], построенную на основе базовой библиотеки LibSVM [Chang, Lin 2011]. Акустические признаки были выбраны на основе набора, использовавшегося в паралингвистическом соревновании ComParE 2013 г. по автоматическому определению социальных сигналов [ComParE 2013; Schuller et al. 2013]. Признаки вычислялись с помощью программной библиотеки openSMILE [Eyben et al. 2010] с окном

обработки 25 мс и шагом 10 мс. Этот набор признаков основывается на 54 низкоуровневых дескрипторах (LLDs): 14 коэффициентов MFCC, логарифм энергии, их производные первого и второго порядка, вероятности озвончения, ЧОТ, частота переходов через ноль и их производные. Для каждого дескриптора считались значения арифметического среднего и стандартного отклонения в сегменте и в восьми ближайших «контекстных» сегментах, эти значения также использовались как признаки. Таким образом, общее количество признаков составило 162 для каждого сегмента аудиосигнала.

Материал, как и в [Verkhodanova et al. 2016], был разделен на два класса: ХЯ и остальная речь. Однако разделение проводилось по файлам и уже на три выборки — обучающую, отладочную и тестовую. Каждый 10-й файл выбирался для обучающего набора, затем снова каждый 10-й — для отладочного набора и остальные — для тестового набора. Подобная процедура повторялась 10 раз.

После обучения классификатора на этапе постобработки применялись фильтр Гаусса и морфологическое открытие (morphological opening). Оба этих метода показали свою эффективность для улучшения значений точности и полноты благодаря учету контекстной информации [Prylipko et al. 2014; Verkhodanova, Shapranov 2016b]. Они применяются для удаления шума при обработке звуковых сигналов и изображений. Фильтр Гаусса используется для сглаживания выбросов, вызванных шумами во входных данных, а морфологическое открытие позволяет отфильтровывать часть ложно-положительных срабатываний, улучшая значение F-меры [Prylipko et al. 2014]. Параметры для фильтра Гаусса и морфологического открытия, а также порог принятия решения, определялись с помощью поиска по сетке параметров на отладочном наборе.

Фильтр Гаусса позволил улучшить значение F-меры на 12 %, а морфологическое открытие — на 2 %. Так, в работе [Verkhodanova, Shapranov 2016a] на корпусе, состоящем из нескольких частей (спонтанные диалоги, записи из мультиязычной речевой базы данных, доклады семинара АРЗ, записи приложения № 5 к БФФ), значение F-меры составило 54 %, а точность и полнота — 55 % и 53 % соответственно. В следующей серии экспериментов с расширенным корпусом, в который дополнительно вошли несбалансированные по полу и специализации диалоги [Verkhodanova et al. 2017], было получено значение F-меры 55 %.

Следующие эксперименты с применением метода опорных векторов производились уже на объединенном корпусе. На данном этапе для обучения модели применялось сглаживание входных данных, при котором отсутствует резкий переход от 0 («не ХЯ») к 1 («ХЯ»): вместо него осуществляется линейный переход (длина перехода — 20 мс), что позволяет снизить влияние переходных участков в начале / конце ХЯ на получившуюся модель. Благодаря этому значение F-меры увеличилось до 56 %.

4. Заключение

В этой статье представлены результаты акустического анализа хезитаций на материале разножанровых записей русской спонтанной и частично подготовленной речи, сделанных в различных условиях и представляющих разнообразные языковые ситуации. Также были описаны различные методы автоматического определения ХЯ в сигнале, которые были опробованы на том же материале. Акустическая картина ХЯ соответствует представлениям о реализации данных явлений в рамках концепции экономии речевых усилий: вокализации имеют тенденцию к централизации в пространстве гласных. Сравнение акустико-просодических маркеров ХЯ в левом и правом контексте показали различия для разных жанров речи. Монологическая речь чаще диалогической содержит явные просодические маркеры наличия хезитации как в левом, так и в правом контексте. Эксперименты по определению ХЯ в русской речи указывают на эффективность и адекватность подходов, не учитывающих лексическую информацию, даже на сложных и зашумленных данных. Лучшим методом

оказался метод опорных векторов, позволивший получить значение F-меры 56 %. Это свидетельствует о возможности использования данного метода для систем АРР, работающих с разнообразными типами речевых сигналов. В будущем использование дополнительной контекстной информации и увеличение базы данных для более сбалансированного разделения на обучающую, отладочную и тестовую выборки может позволить применить дополнительные техники сглаживания, уменьшить количество ложных срабатываний и таким образом улучшить результат.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

АРЗ 2011 — «Анализ разговорной русской речи» АРЗ: Труды пятого междисциплинарного семинара. СПб.: ГУАП, 2011. [*«Analiz razgovornoi russkoi rechi» AR3: Trudy pyatogo mezhdistsiplinarnogo seminara.* [“Russian colloquial speech analysis” AR3: Proceedings of the 5th interdisciplinary seminar.] St. Petersburg: GUAP, 2011.]

Богданова-Бегларян 2014 — Богданова-Бегларян Н. В. Прагматемы в устной повседневной речи: определение понятия и общая типология // Вестник Пермского университета. Российская и зарубежная филология. 2014. № 3. С. 7–20. [Bogdanova-Beglaryan N. V. Pragmatemes in oral colloquial speech: Definition and general typology. *Vestnik Permskogo universiteta. Rossiiskaya i zarubezhnaya filologiya.* 2014. No. 3. Pp. 7–20.]

Богданова-Бегларян 2016 — Богданова-Бегларян Н. В. Вербальные хезитативы русской устной речи: реализация поисковой функции и «рефлекс поиска» // Язык и метод: Русский язык в лингвистических исследованиях XXI века. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego, 2016. С. 345–354. [Bogdanova-Beglaryan N. V. Verbal hesitatives in spoken Russian: realization of the searching function and “the search reflex”. *Yazyk i metod: Russkii yazyk v lingvisticheskikh issledovaniyakh XXI veka.* Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego, 2016. Pp. 345–354.]

Верходанова 2013 — Верходанова В. О. Алгоритмы и программные средства автоматического определения речевых сбоев в звуковом сигнале // Труды СПИИРАН. 2013. № 31. С. 43–60. [Verkhodanova V. O. Algorithms and software for automatical speech disfluency detection in audio signal. *Trudy SPIIRAN*. 2013. No. 31. Pp. 43–60.]

Кибрик, Подлесская 2014 — Кибрик А. А., Подлесская В. И. Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: Litres, 2014. [Kibrik A. A., Podlesskaya V. I. *Rasskazy o snovideniyakh: Korpusnoe issledovanie ustnogo russkogo diskursa.* [The dream stories: A corpus study of spoken Russian discourse.] Moscow: Litres, 2014.]

Кипяткова, Карпов 2016 — Кипяткова И. С., Карпов А. А. Разновидности глубоких искусственных нейронных сетей для систем распознавания речи // Труды СПИИРАН. 2016. № 6. С. 80–103. [Kipyatkova I. S., Karpov A. A. Deep artificial neural network types for speech recognition systems. *Trudy SPIIRAN.* 2016. No. 6. Pp. 80–103.]

Подлесская, Кибрик 2007 — Подлесская В. И., Кибрик А. А. Самоисправления говорящего и другие типы речевых сбоев как объект аннотирования в корпусах устной речи // Научно-техническая информация. Серия 2. № 2. 2007. С. 2–23. [Podlesskaya V. I., Kibrik A. A. Speaker’s self-corrections and other types of disfluency as an object of annotation in spoken language corpora. *Nauchno-tekhnicheskaya informatsiya.* Series 2. No. 2. 2007. Pp. 2–23.]

Хуршудян 2005 — Хуршудян В. Экспериментальное исследование хезитации в разноструктурных языках // Труды конференции “Dialog’2005”. 2005. С. 497–501. [Khurshudyan V. Experimental study of hesitations in languages of different structures. *Trudy konferentsii “Dialog’2005”.* 2005. Pp. 497–501.]

Akusok et al. 2015 — Akusok A., Björk K.-M., Miche Y., Lendasse A. High-performance extreme learning machines: A complete toolbox for big data applications. *IEEE Access.* 2015. Vol. 3. Pp. 1011–1025.

Arbisi-Kelm, Jun 2005 — Arbisi-Kelm T., Jun S. A. A comparison of disfluency patterns in normal and stuttered speech. *Disfluency in Spontaneous Speech*. 2005. Pp. 13–16.

Audhkhasi et al. 2009 — Audhkhasi K., Kandhway K., Deshmukh O. D., Verma A. Formant-based technique for automatic filled-pause detection in spontaneous spoken English. *Proc. of the ICASSP-2009*. 2009. Pp. 4857–4860.

Barnes 2006 — Barnes J. *Strength and weakness at the interface: Positional neutralization in phonetics and phonology.* Berlin: Walter de Gruyter, 2006.

Boersma, Weenink 2016 — Boersma P., Weenink D. *Praat: doing phonetics by computer [computer program], version 6.0.11*. Available at: http://www.praat.org/

Chafe 1980 — Chafe W. L. (ed.). *The pear stories: cognitive, cultural, and linguistic aspects of narrative production*. Norwood (Mass.): Ablex Publishing Corp, 1980.

Chang, Lin 2011 — Chang C. C., Lin C. J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2011. Vol. 2. Pp. 1–127.

ComParE 2013 — *INTERSPEECH: Computational Paralinguistic Challenge, 2013*. Available at: http://emotion-research.net/sigs/speech-sig/is13-compare.

DiSS'03 2003 — Proceedings of DiSS'03, disfluency in spontaneous speech workshop. *Papers in Theoretical Linguistics 90*, Sweden, Göteborg University, 2003. Pp. 3–4.

Eisler 1968 — Eisler F. G. *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press, 1968.

English CTS — *LDC: English CTS treebank with structural metadata*. Available at: http://catalog.ldc.upenn.edu/LDC2009T01

Esposito et al. 2016 — Esposito A., Esposito A. M., Likforman-Sulem L., Maldonato M. N., Vinciarelli A. On the significance of speech pauses in depressive disorders: Results on read and spontaneous narratives. *Recent Advances in Nonlinear Speech Processing*. 2016. Pp. 73–82.

Eyben et al. 2010 — Eyben F., Wöllmer M., Schuller B. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. *Proc. Of the Multimedia ACM Multimedia 2010*, *Firenze, Italy*. 2010. Pp. 1459–1462.

Ferreira et al. 2004 — Ferreira F., Lau E F., Bailey K. G. D. Disfluencies, language comprehension, and tree adjoining grammars. *Cognitive Science*. 2004. Vol. 28. No. 5. Pp. 721–749.

Garg, Ward 2006 — Garg G., Ward N. Detecting filled pauses in tutorial dialogs. *Departmental Technical Reports (CS). 2006. Paper 199*. Available at: http://digitalcommons.utep.edu/cs_techrep/199

Giannini 2003 — Giannini A. Hesitation phenomena in spontaneous Italian. *Proc. Of the ICPhS-2003, Barcelona, Spain*. 2003. Pp. 2653–2656.

Godfrey et al. 1992 — Godfrey J. J., Holliman E. C., McDaniel J. Switch board: Telephone speech corpus for research and development. *Proc. of the ICASSP-1992, San Francisco, USA*. 1992. Vol. 1. Pp. 517–520.

Goto et al. 1999 — Goto M., Itou K., Hayamizu S. A real-time filled pause detection system for spontaneous speech recognition. *Proc. of the Eurospeech-1999, Budapest, Hungary*. 1999. Pp. 227–230.

Gupta et al. 2013 — Gupta R., Audhkhasi K., Lee S., Narayanan S. Paralinguistic event detection from speech using probabilistic time-series smoothing and masking. *Proc. of the INTERSPEECH-2013, Lyon, France*. 2013. Pp. 173–177.

Kaya et al. 2017 — Kaya H., Salah A., Karpov A., Frolova O., Grigorev A., Lyakso E. Emotion, age, and gender classification in children's speech by humans and machines. *Computer Speech and Language*. 2017. Vol. 46. Pp. 268–283.

Kaya, Karpov 2018 — Kaya H., Karpov A. Efficient and effective feature normalization strategies for cross-corpus acoustic emotion recognition. *Neurocomputing*. 2018. Vol. 275. Pp. 1028–1034.

Liu 2004 — Liu Y. *Structural event detection for rich transcription of speech.* Ph.D thesis. Purdue University, 2004.

Liu et al. 2006 — Liu Y., Shriberg E., Stolcke A., Hillard D., Ostendorf M., Harper M. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech and Language Processing*. 2006. Vol. 14. No. 5. Pp. 1526–1539.

Medeiros et al. 2013a — Medeiros H., Moniz H., Batista F., Trancoso I., Nunes L. Disfluency detection based on prosodic features for university lectures. *Proc. of the INTERSPEECH-2013, Lyon, France.* 2013. Pp. 2629–2633.

Medeiros et al. 2013b — Medeiros H., Batista F., Moniz H., Trancoso I., Meinedo H. Experiments on automatic detection of filled pauses using prosodic features. *Actas de Inforum*. 2013. Pp. 335–345.

Moniz et al. 2014 — Moniz H., Batista F., Mata A. I., Trancoso I. Speaking style effects in the production of disfluencies. *Speech Communication.* 2014. Vol 65. Pp. 20–35.

O'Connell, Kowal 2004 — O'Connell D. C., Kowal S. The history of research on the filled pause as evidence of the written language bias in linguistics (Linell, 1982). *Journal of Psycholinguistic Research*. 2004. Vol. 33. No. 6. Pp. 459–474.

Ogden 2001 — Ogden R. Turn-holding, turn-yielding and laryngeal activity in Finnish talking-interaction. *Journal of the International Phonetics Association*. 2001. Vol. 31. No. 1. Pp. 139–152.

O'Shaughnessy 1992 — O'Shaughnessy D. Recognition of hesitations in spontaneous speech. *Proc. of the ICASSP'92*. 1992. Vol. 1. Pp. 521–524.

Prylipko et al. 2014 — Prylipko D., Egorow O., Siegert I., Wendemuth A. Application of image processing methods to filled pauses detection from spontaneous speech. *Proc. of the INTERSPEECH-2014, Singapore*. 2014. Pp. 1816–1820.

Schuller et al. 2013 — Schuller B., Steidl S., Batliner A., Vinciarelli A., Scherer K., Ringeval F., Chetouani M., Weninger F., Eyben F., Marchi E., Mortillaro M., Salamin H., Polychroniou A., Valente F., Kim S. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism. *Proc. of the INTERSPEECH-2013, Lyon, France*. 2013. Pp. 148–152.

Scikit-Learn — *Scikit-Learn: Machine learning in Python*. Available at: http://scikit-learn.org.

Shriberg 1994 — Shriberg E. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis. Univ. of California at Berkeley, 1994.

Shriberg 2001 — Shriberg E. To 'Errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*. 2001. Vol. 31. No. 1. Pp. 153–169.

Shriberg 2005 — Shriberg E. Spontaneous speech: How people really talk and why engineers should care. *Proc. of the INTERSPEECH-2005, ISCA, Lisbon, Portugal*. 2005. Pp. 1781–1784.

Shriberg et al. 1997 — Shriberg E., Bates R. A., Stolcke A. A prosody only decision-tree model for disfluency detection. *Proc. of the EUROSPEECH-1997, Rhodes, Greece*. 1997. Pp. 2383–2386.

Skrelin et al. 2010 — Skrelin P., Volskaya N., Kocharov D. et al. A fully annotated corpus of Russian speech. *Proc. of the LREC'10, Valletta, Malta*. 2010. Pp. 109–112.

Snyman 2005 — Snyman J. Practical mathematical optimization: An introduction to basic optimization theory and classical and new gradient-based algorithms. Vol. 97. *Springer Science & Business Media*. 2005.

Stepanova 2007 — Stepanova S. Some features of filled hesitation pauses in spontaneous Russian. *Proc. of the ICPhS-2007, Saarbrucken, Germany*. 2007. Vol. 16. Pp. 1325–1328.

Stolcke et al. 1998 — Stolcke A., Shriberg E., Bates R. A. et al. Automatic detection of sentence boundaries and disfluencies based on recognized words. *Proc. of the ICSLP-1998*. 1998. Vol. 2. Pp. 2247–2250.

Stouten, Martens 2003 — Stouten F., Martens J. P. A feature-based filled pause detection system for Dutch. *Proc. of the ASRU'03, IEEE*. 2003. Pp. 309–314.

Verkhodanova, Shapranov 2015 — Verkhodanova V., Shapranov V. Multi-factor method for detection of filled pauses and lengthenings in Russian spontaneous speech. *Proc. of the SPECOM-2015*. 2015. Pp. 285–292.

Verkhodanova, Shapranov 2016a — Verkhodanova V., Shapranov V. Detecting filled pauses and lengthenings in Russian spontaneous speech using SVM. *Proc. of the SPECOM-2016, Budapest, Hungary. Lecture Notes in Computer Science*. 2016. Vol. 9811. Pp. 224–231.

Verkhodanova, Shapranov 2016b — Verkhodanova V., Shapranov V. Experiments on detection of voiced hesitations in Russian spontaneous speech. *Journal of Electrical and Computer Engineering*. 2016. Available at: http://dx.doi.org/10.1155/2016/2013658

Verkhodanova et al. 2016 — Verkhodanova V., Shapranov V., Karpov A. Filled pauses and lengthenings detection using machine learning techniques. *Proc. of the ExLing, Saint Petersburg, Russia*. 2016. Pp. 175–178.

Verkhodanova et al. 2017 — Verkhodanova V., Shapranov V., Kipyatkova I. Hesitations in spontaneous speech: Acoustic analysis and detection. *Proc. of the SPECOM-2017, Hatfield, UK*. 2017. Pp. 398–406.

Zahorian et al. 2011 — Zahorian S. A., Wu J., Karnjanadecha M., Sekhar Vootkuri C., Wong B., Hwang A., Tokhtamyshev E. Open source multi-language audio database for spoken language processing applications. *Proc. of the INTERSPEECH-2011*. 2011. Pp. 1493–1496.

Получено / received 07.11.2017 *Принято / accepted 14.06.2018*