УДК 004.522

В.О. Верходанова

Разработка средств автоматического определения речевых сбоев в спонтанной русской речи

Предложен алгоритм автоматического определения наиболее частотных речевых сбоев и артефактов в спонтанной русской речи, основанный на анализе акустических параметров. При выделении звонких хезитационных явлений использовался кросскорреляционный метод для выделения глухих хезитационных явлений и дыхания – метод полосовой фильтрации. Эксперименты проводились на специально собранном корпусе спонтанной русской речи, состоящем из 18 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.

Проблема речевых сбоев и методы их исследования. В настоящее время автоматическое определение речевых сбоев (например, пауз хезитаций, самокоррекций) и артефактов (смех, вздох) актуально как для обработки звучащей речи, так и ее транскрипций, поскольку они являются источником ошибок при распознавании, что приводит к ошибкам в транскрипциях [1–3]. Для исследования этих явлений используются корпусы спонтанной речи с многоуровневой аннотацией. Для того чтобы в корпусе помимо такой информации, как фонемы, слова, синтагмы, дополнительно отмечались речевые сбои, используется Rich Transcription – транскрипция, в которой учитываются границы предложений, слова-заполнители, речевые сбои и др. [4]. Данная разметка применялась при обработке корпуса, разработанного на кафедре фонетики СПбГУ. Этот корпус включает в себя речь 4 мужчин и 4 женщин и учитывает различные произносительные стили. Аннотация производилась на 6 уровнях, где отмечалась вся фонетическая и просодическая информация о записанной речи [5]. Среди зарубежных корпусов широко известен корпус SWITCHBOARD [6], который представляет собой более 240 ч записи спонтанных телефонных разговоров более 500 дикторов обоих полов. Речь полностью затранскрибирована, и транскрипции проверены автоматически и дважды вручную. Всего было размечено 78 типов неречевых данных, включая вздох, кашель, зевание, мяуканье, звон посуды и др.

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 посвящено моделированию речевых сбоев в рамках создания систем автоматического распознавания речи [7–9]. 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 [10, 11] или устранения сбоев с использованием транскрипций речи [7]. Методы параметрической обработки речевых сбоев опираются только на акустические параметры [8, 12], тогда как комбинированные модели используют дополнительно языковое моделирование [7]. Для применения языковой модели дополнительно необходим большой корпус транскрипций, по которому происходит обучение модели, в то время как параметрическая обработка не требует транскрибирования сигнала [13]. В данном исследовании было решено опираться на акустические характеристики речевых сбоев и артефактов.

Описание собранного корпуса спонтанной диалогической русской речи. Материалом для исследования послужил собранный корпус 18 спонтанных диалогов на русском языке. Диалоги представляли собой выполнение собеседниками двух задач: описание маршрута по карте и нахождение общего свободного времени по расписанию. Подобная методика записи корпуса использовалась для того, чтобы максимально приблизить записываемую речь к ее разговорной форме. Соглас-

но Е.А. Земской [14], основными условиями непринужденного неподготовленного непосредственного общения является неофициальность отношений между собеседниками, их непосредственное участие в разговоре, устная форма, опора на внеязыковую ситуацию, использование в основном жанра диалога. Собранный корпус представляет собой 50 мин диалогической речи – 18 диалогов длительностью от 1,5 до 5 мин. Запись проводилась в звукоизолированной комнате с использованием современных мобильных планшетов Samsung Galaxy Tab 2 приложением Smart Voice Recorder. Записанные аудиофайлы имеют следующие параметры: частота дискретизации 16 кГц, битрейт – 256 кбит/с, количество каналов – 1. Все записи были сделаны в Санкт-Петербурге в конце 2012 – начале 2013 г. 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 не оконченным высшим образованием. Участвовавшие в записи 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 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 – заполненные паузы типа [ə] или [ɐ]; в класс удлинений (хезитационных удлинений) – 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 рис. 1.

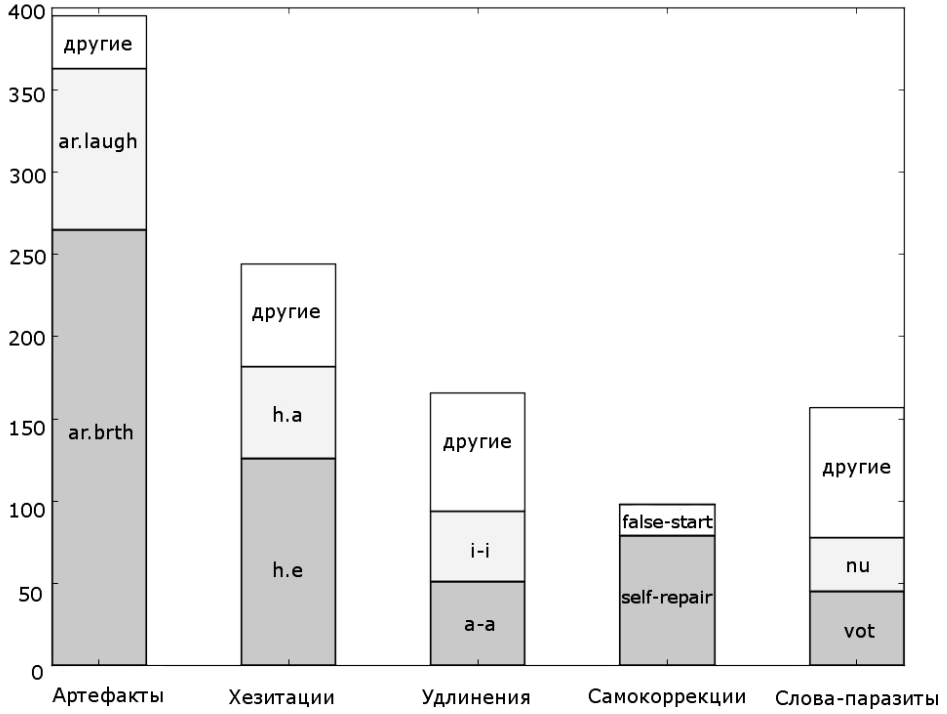

Рис. 1. Распределение элементов по классам (где ar.laugh – смех; ar.brth – вздохи; h.a – хезитация [ɐ]; h.e – хезитация [ə]; i-i – хезитационное удлинение /i/; a-a – хезитационное удлинение /a/; false-start – фальстарт; self-repair – самокоррекция; nu – «ну», vot – «вот»)

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».

Алгоритм автоматического определения хезитаций в речевом 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), постоянство спектра и длительность, превышающая 150–200 мс. Хезитационным явлениям свойственно постоянство ЧОТ, но не все интервалы, где ЧОТ постоянна, можно отнести к этим явлениям. Например, слова «мне», «невменяемый», «налево» будут иметь длинный участок непрерывной и постоянной ЧОТ. 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 [15]. Поиск звонких хезитационных явлений (пауз хезитации и хезитационных удлинений) производился на основе алгоритма, показанного на рис. 2.

Рис. 2. 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 (0,18 с): такие интервалы считались соответствующими искомым удлинениям.

Что касается рассматриваемых в данном исследовании артефактов, были взяты наиболее распространенные из них: вздохи и громкое дыхание. Поиск этих артефактов, а также удлиненных глухих щелевых согласных производился на основе алгоритма, показанного на рис. 3. Такие элементы – это шум разной громкости и распределения мощности по спектру. Поэтому для таких явлений амплитуда сигнала больше, чем на участках с «тишиной», и на спектрограмме нет участков с постоянными значениями формант. А глухие щелевые согласные имеют характерные полосы в спектре. Например, свистящие согласные имеют максимальную интенсивность спектра на частоте 4500–7000 Гц. 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 сигнал/шум 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 среднего значения спектра внутри каждой из полос, заданных для глухих щелевых согласных, к среднему значению по всему спектру, порог равен 2 (также определен экспериментально) [15]. 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.

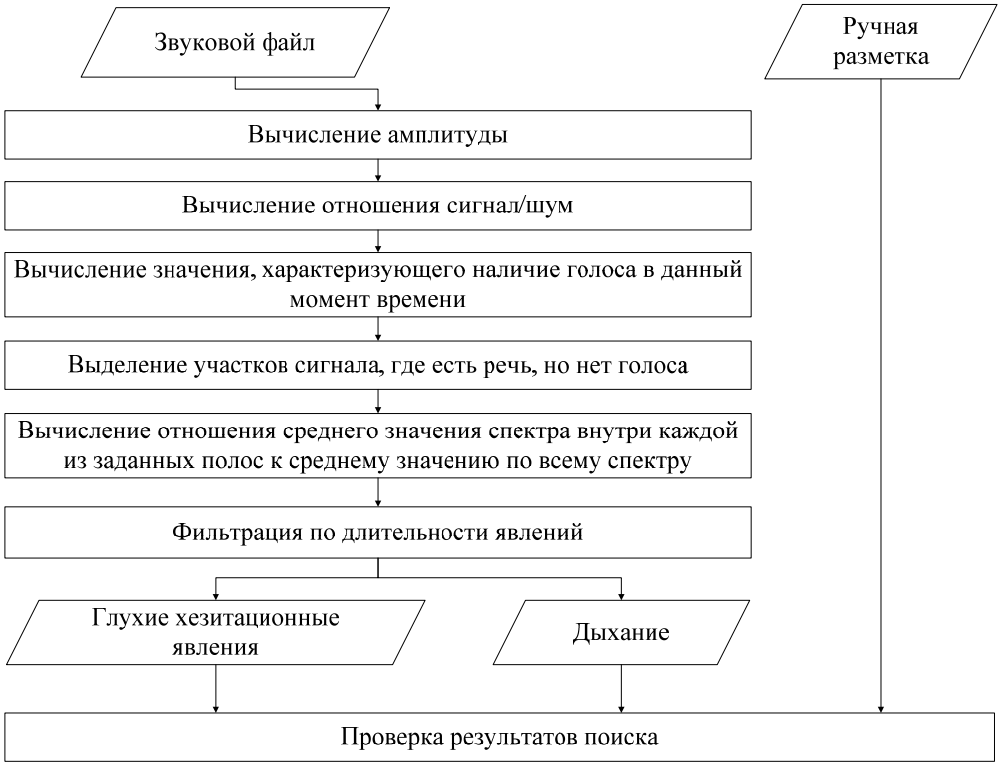

Рис. 3. Схема поиска звонких хезитационных явлений

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% . 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. Таким образом, если решить проблему с ларингализованными звуками, то доля правильного определения вырастет до 88%. Точность определения артефактов составила 66%, если пренебречь тихим дыханием, которое составило больше половины ненайденных явлений, и причмокиваниями, нахождение которых не предполагалось в связи с ограничением на длительность элементов (минимум 200 мс).

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

Данное исследование поддержано фондом РФФИ (проект № 12-06-31203-мол_а).

Литература

1. Верходанова В.О. Моделирование речевых сбоев в системах автоматического распознавания речи / В.О. Верходанова, А.А. Карпов // Вестник Том. гос. ун-та. – 2012. – № 363. – С. 10–15.

2. Ронжин А.Л. Особенности дистанционной записи и обработки речи в автоматах самообслуживания / А.Л. Ронжин, А.А. Карпов, И.А. Кагиров // Информационно-управляющие системы. – 2009. – Вып. 42, т. 5. – С. 32–38.

3. Бондаренко В.П. Сегментация и параметрическое описание речевого сигнала / В.П. Бондаренко, А.А. Конев, Р.В. Мещеряков // Изв. высш. учеб. заведений. Приборостроение. – 2007. – Т. 50, № 10. – С. 3–7.

4. Liu Y. Structural Event Detection for Rich Transcription of Speech: PhD thesis. – Berkeley: Purdue University and ICSI, 2004. – 253 p.

5. Skrelin P. A Fully Annotated Corpus of Russian Speech / P. Skrelin, N. Volskaya, D. Kocharov et al. // In Proceedings of the 7-th Conference on International Language Resources and Evaluation LREC'10 (Valletta, Malta). – 2010. – P. 109–112.

6. SWITCHBOARD: A User's Manual. URL [Электронный ресурс]. – Режим доступа: http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html (дата обращения: 05.09.2013).

7. Karpov A. Large vocabulary Russian speech recognition using syntactico-statistical language modeling / A. Karpov, K. Markov, I. Kipyatkova et al. // Speech Communication. – 2013. – Vol. 56. – P. 213–228.

8. Liu Y. Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies / Y. Liu, E. Shriberg, A. Stolcke et al. // IEEE Transactions on Audio, Speech and Language Processing. – 2006. – № 14(5). – P. 1526–1540.

9. Ронжин А.Л. Проектирование интерактивных приложений с многомодальным интерфейсом / А.Л. Ронжин, А.А. Карпов // Доклады ТУСУРа. – 2010. – № 1 (21), ч. 1. – С. 124–127.

10. Kaushik M. Automatic Detection and Removal of Disfluencies from Spontaneous Speech / M. Kaushik, M. Trinkle, A. Hashemi-Sakhtsari // In Proceedings of the 13-th Australasian International Conference on Speech Science and Technology SST (Melbourne, Australia). – 2010. – P. 98–101.

11. Сизов А.Г. Оценка качества голоса лиц, страдающих дисфонией / А.Г. Сизов, Р.В. Мещеряков, С.Д. Тиунов // Труды СПИИРАН. – 2012. – № 1. – С. 138–152.

12. Мещеряков Р.В. Система оценки качества передаваемой речи // Доклады Том. госу. Ун-та систем управления и радиоэлектроники. – 2010. – № 2(22), ч. 1. – С. 324–329.

13. Кипяткова И.С. Сегментация паралингвистических фонационных явлений в спонтанной русской речи / И.С. Кипяткова, В.О. Верходанова, А.Л. Ронжин // Вестник Перм. ун-та. Российская и зарубежная филология. – 2012. – Вып. 2 (18). – С. 17–23.

14. Земская Е.А. Русская разговорная речь: лингвистический анализ и проблемы обучения. – М.: Русский язык, 1987. – 240 с.

15. Verkhodanova V. Automatic Detection of Speech Disfluencies in the Spontaneous Russian Speech / V. Verkhodanova, V. Shapranov // Springer International Publishing Switzerland / M. Zelezny et al. (Eds.): SPECOM 2013, LNAI 8113. – 2013. – P. 70–77 [Электронный ресурс]. – Режим доступа: http://link.springer.com/chapter/10.1007%2F978-3-319-01931-4_10, свободный (дата обращения: 02.10.2013).

Верходанова Василиса Олеговна
Мл. науч. сотрудник лаб. речевых и многомодальных интерфейсов СПИИРАН, Санкт-Петербург
Тел.: 8 (812) 3-28-70-81
Эл. почта: verkhodanova@iias.spb.su, interiora@gmail.com

Verkhodanova V.O.

Development of automatic detection method for speech disfluencies in spontaneous Russian speech

An algorithm for automatic detection of most frequent speech disfluencies and artefacts in the spontaneous Russian speech was proposed basing on the analysis of acoustical features of elements in question. 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 especially collected corpus of 18 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.