

Фонетическое транскрибирование речевых корпусов: проблемы и решения

В наше время появляется все больше звуковых корпусов русской речи. Явление это очень отрадное, хотя далеко не все из них доступны для широкого круга пользователей. Обычно звуковые файлы таких корпусов сопровождаются текстовыми файлами с орфографической транскрипцией, но акустико-фонетическая транскрипция того, что реально прозвучало в сигнале, как правило, отсутствует. Последнее существенно снижает ценность получаемых данных.

Между тем, точная информация о реальном фонетическом содержании звучащего речевого сигнала могла бы быть интересна по крайней мере трем группам исследователей.

Для лингвистики создание корпусов устной речи очевидным образом означает прежде всего встречу с ее подлинным объектом – естественным звучащим дискурсом, и последствия этой встречи не могут в перспективе не привести к уточнению, расширению и, возможно, пересмотру сложившихся представлений о языке, и отнюдь не только в области фонетики и фонологии.

Что касается систем автоматического распознавания речи (АРР), то слова в них обычно распознаются на базе вероятностного сравнения акустических признаков отрезка речевого сигнала с акустическими моделями соответствующих единиц, хранящимися в «словаре» возможных реализаций словоформ языка. Наполнение такого «словаря» вариантами «естественных» реализаций происходит в основном путем трансформирования по каким-то правилам теоретических («базовых») фонемных транскрипций орфографического представления соответствующих словоформ (Кипяткова 2008). Наличие речевых корпусов достаточного объема, снабженных акустико-фонетической транскрипцией, позволило бы точнее сформулировать подобные правила или использовать для получения акустических моделей саму транскрипцию, что в перспективе повысило бы надежность автоматического распознавания речи.

Не менее полезной оказалась бы информация о реальной акустической реализации словоформ в естественной речи и для создателей систем синтеза речи по тексту (Лобанов, Цирульник 2007).

В моделировании процессов восприятия речи человеком первостепенная задача – преобразование непрерывного речевого акустического сигнала в последовательность лексических единиц, т.е. членение потока на дискретные отрезки и лексическая идентификация таких отрезков. В экспериментах с лишенными пробелов письменными текстами (Kassevich et al. 2000) была проверена возможность сегментации через идентификацию методом «когорты»

с обращением к словарю словоформ. Количество ошибочных членений и отказов не превышало 2–2,5 процентов.

В отличие от письменного текста, для естественного речевого сигнала, и особенно спонтанной речи, характерна систематическая редукция фонетического облика словоформ. Возникновение потребности в речевых корпусах большого объема, снабженных акустико-фонетической транскрипцией, не в последнюю очередь объясняется необходимостью получения надежной информации о том, какой именно сигнал поступает на вход системы восприятия речи человеком.

Имеющийся у нас скромный опыт сплошного акустико-фонетического транскрибирования фрагментов спонтанной речи общей длительностью около 90 минут (Венцов, Слепокурова 2010; Венцов и др. 2011; Венцов, Слепокурова 2012) заставляет по-новому взглянуть на такие проблемы, как перцептивная сегментация речевого сигнала, лексический поиск, а также структура перцептивного словаря. Ниже приведены характерные примеры полученной нами транскрипции (весь материал – на сайте www.narusco.ru).

вдох	0,738
как_говорится, [kыgəˈr'i+cə]	0,534
дорогу_осилит идущий, [dəro+veˈs'il'i+t eduˈ+šč]	1,172
и вот именно с этими детьми [i vot i+m'əna s e+t'im'ie:: d'it'm'i+]	1,740
кхэ	0,264

вдох	0,502
американским она понравилась [aˈm'ir'ika+nsk'im ana+ panre+els]	1,271
потому_что_они ничего не читают [ptamu+štaˈn'i+ n'ičö+ n'i čita+et]	1,863
вдох	0,403

пауза	0,313
я думаю никто с этим не [e du+m n'ikto+ s e+t'im n'e]	1,046
пауза	0,316
будет_спорить [bu+d'ˈespoˈr'it']	0,966
пауза	0,448
но [no+]	0,216
пауза	0,123

Даже из этих кратких примеров видно, что в спонтанной речи наблюдается стяжение гласных и согласных на стыках словоформ, масштабная редукция заударных компонентов словоформ, реализация гласных, качество которых даже в позиции под ударением отличается от предписанного правилами орфоэпии. Кроме того, выпадение слогов и перцептивное выделение безударных в норме слогов заставляет усомниться в важности для восприятия информации о точном фонемном качестве гласных, а также постулируемой в некоторых работах роли ритмической структуры словоформы (фразы).

Наконец, создается впечатление, что в перцептивном словаре носителя языка могли бы содержаться и часто встречающиеся редуцированные варианты словоформ: *щас* (сейчас), *те* (тебе), *тока* (только) и т.п.

Можно отметить, что вышеперечисленные явления, судя по имеющимся данным, не в меньшей степени характерны и для английской спонтанной речи (Dilley, Pitt 2007; Warner 2012).

Естественно, что создание доступных речевых корпусов, снабженных тщательной акустико-фонетической транскрипцией, существенно расширило бы наши знания об особенностях естественного речевого сигнала и сделало бы более продуктивными многие фундаментальные лингвистические исследования и прикладные технические разработки.

Беда, однако, в том, что процесс получения подобной транскрипции из-за отсутствия адекватных автоматических систем требует огромных затрат времени квалифицированных экспертов, и именно это, вероятно, и является главной причиной отсутствия полноценных корпусов устной речи.

Попытки использования для этой цели действующих систем АРР, насколько можно судить по публикациям, являются несостоятельными, поскольку из-за большого количества ошибок приводят в результате к необходимости последующей ручной правки (Bertrand et al. 2006; Schuppler 2011). При этом ошибочно определенными оказываются не только границы выделяемых сегментов, но и их фонетические признаки, поскольку для целей автоматического распознавания акустические модели создаются применительно к каноническому фонемному представлению словоформ. В результате подобного автоматического транскрибирования позиционный аллофон описывается символом той фонемы, которая должна была бы быть реализована в данной позиции (рис. 1). И это при условии, что в этом конкретном случае «распознавание» производилось с использованием словаря трифонов, а не слов.

Действительно, как следует из динамической спектрограммы на рис. 1, второй гласный в слове /amerika/ был произнесен диктором как [i], но при автоматическом транскрибировании он был заменен «нужной» фонемой, поскольку систему распознавания «научили» именно этому.

Видимо, именно из-за появления такого рода ошибок в серьезных лингвистических исследованиях, претендующих на выявление и интерпретацию тонких фонетико-фонологических особенностей звучащей речи, используются речевые корпуса, полностью транскрибированные вручную (Dilley, Pitt 2007)

Представляется, что проблема заключается не столько собственно в алгоритмах распознавания, заложенных в основу современных систем АРР, сколько в структуре и способах описания единиц используемых при этом «словарей». Возможно, выходом могло бы стать использование в системах автоматического транскрибирования, к примеру, словарей открытых слогов или трифонов, описываемых последовательностью не фонем, а позиционных аллофонов, а также создание соответствующих акустических моделей. Процесс создания и пополнения словаря в такой системе мог бы быть итеративным:

начальная версия создается на основании предварительно транскрибированного экспертами звучащего текста небольшого объема, следующий текст транскрибируется автоматически, затем вручную исправляются ошибки, словарь пополняется и т.д.

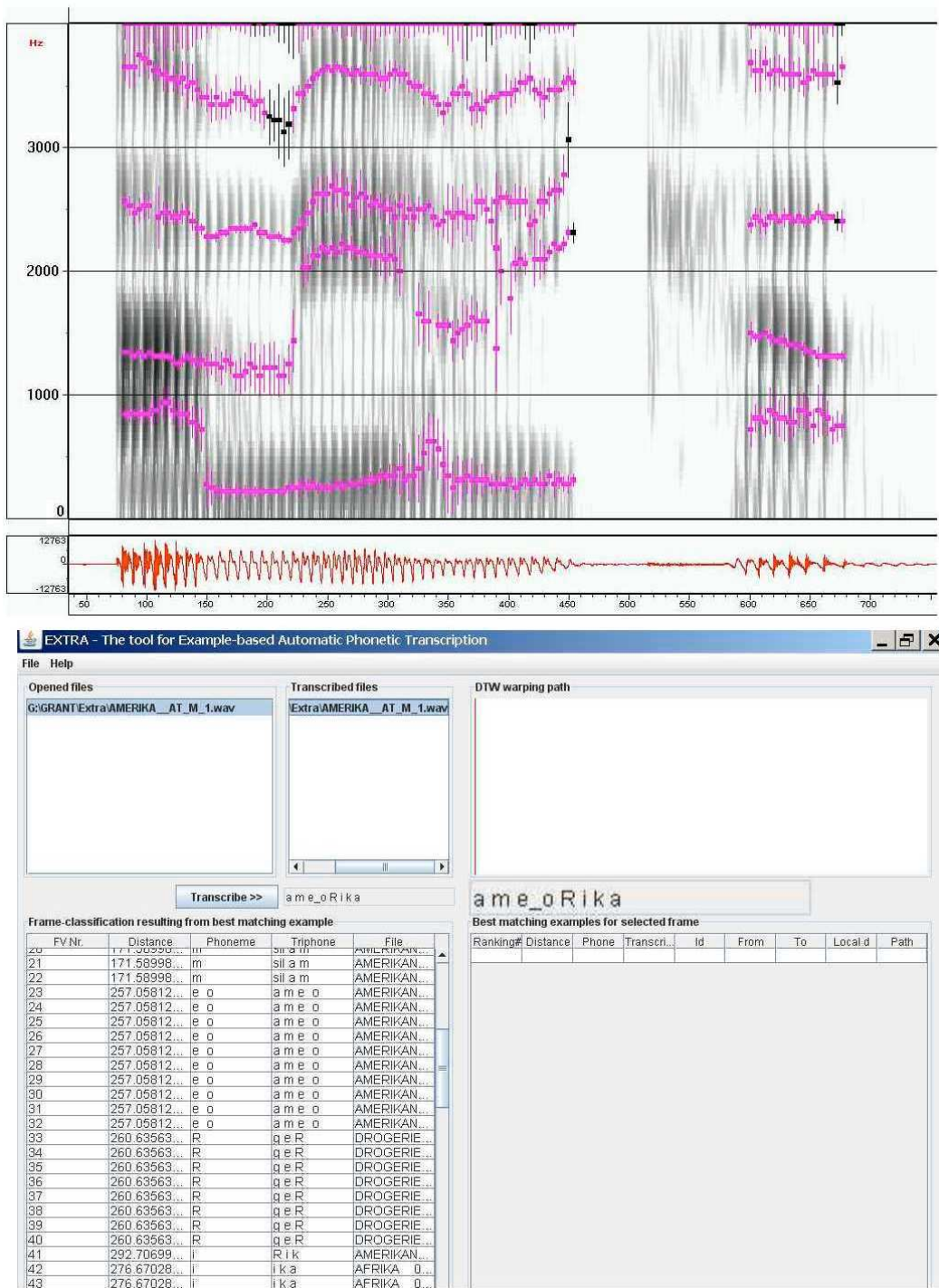


Рис. 1. Спектрограмма слова *amerika* из демонстрационной версии программы Extra (Leitner et al. 2010) и результат его автоматического транскрибирования данной программой.

Однако следует иметь в виду, что и при транскрибировании вручную возможно и, более того, неизбежно появление значительного числа ошибок, если процедурой не предусмотрена минимизация влияния на принимаемые экспертами решения знаний о языке.

Иногда в инструкциях экспертам, выполняющим транскрибирование речевого корпуса, предлагается, как ни странно, в большей степени ориентироваться на лексическое значение описываемого отрезка речевого сигнала, чем на его истинное звучание. Рекомендуются, в частности, прослушивать речевой сигнал порциями длительностью в 0,5-1 секунду (Gillis 2001). В такой интервал может уложиться несколько знаменательных слов, и велика опасность того, что эксперту при оценке реального звучания будет психологически трудно абстрагироваться от лексического образа услышанных слов и знания правил их произнесения. Как известно, слушаем мы ушами, а слышим головой (мозгом).

Наш опыт транскрибирования свидетельствует, что избежать влияния лингвистических знаний эксперта удастся, если ограничить длительность единовременно анализируемого (прослушиваемого) сигнала последовательностью согласный-гласный или гласный-согласный, а иногда и вообще всего одним звуком. При этом в обязательном порядке следует использовать возможности, предоставляемые синхронным просмотром динамических спектрограмм, позволяющим в большинстве случаев без труда устанавливать границы отдельных сегментов.

Анализ динамических спектрограмм с одновременным прослушиванием выделенного отрезка сигнала позволяет надежно идентифицировать подавляющее число согласных, причем независимо от особенностей дикторского произнесения. Сложнее идентифицировать гласные, особенно с учетом сдвига их формантных частот у женских голосов. Для идентификации гласных мужских голосов большим подспорьем для нас стали графики рассеивания частот двух первых формант, полученные В.Б. Кузнецовым (2004). При этом оказывается, что результаты экспертных оценок по частотам формант совпадают с аудиторскими оценками соответствующих звуков (рис. 2).

Таким образом, выяснилось, что достаточно надежная идентификация качества гласных может быть достигнута почти исключительно на основе анализа траекторий частот формант и их сопоставления с типичными распределениями этих частот для разных групп дикторов, особенно если сведения о мужских голосах, полученные В.Б. Кузнецовым, дополнить аналогичными данными для высоких женских голосов, которые можно накапливать уже в процессе транскрибирования.

При наличии согласованных с разными группами исследователей правил транскрибирования можно было бы организовать процесс обработки имеющихся в свободном доступе звучащих текстов, например, силами волонтеров, используя возможности Интернета, как это уже делают организаторы «Открытого корпуса» для морфологического аннотирования текстов.

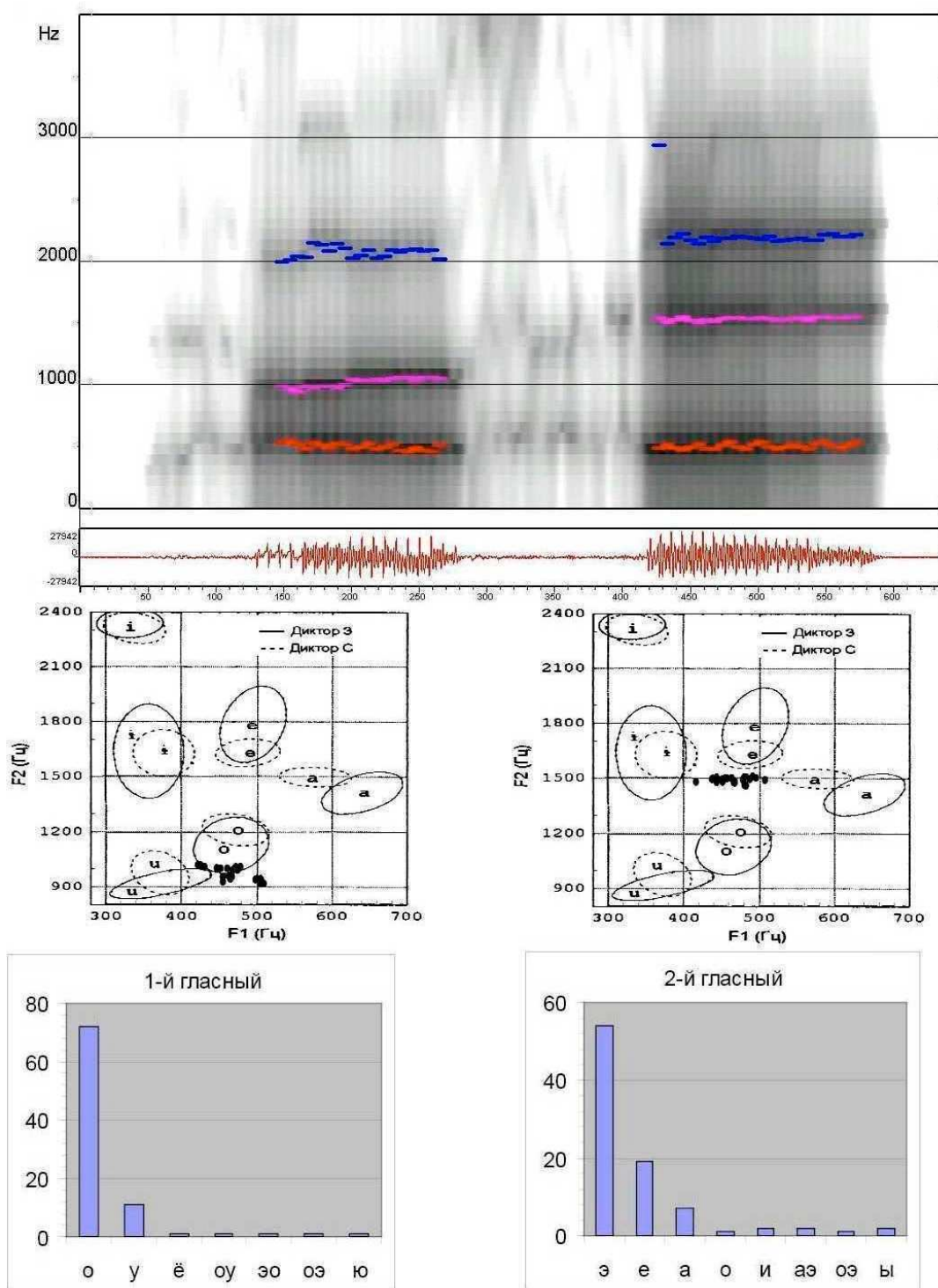


Рис. 2. Динамическая спектрограмма, частоты формант гласных и гистограммы числа аудиторских оценок двуслога ПОСКО из фразы «как бы поскорей удрать из этой школы». По: Апушкина 2011.

Литература

- Апушкина И.Е. Безударные гласные в спонтанном тексте // Проблемы социо- и психолингвистики: Выпуск 15: Пермская социопсихолингвистическая школа: идеи трех поколений: К 70-летию Аллы Соломоновны Штерн / Отв. ред. Е.В.Ерофеева.– Пермь: [б.и.], 2011.– С. 38–45.
- Венцов А.В., Слепокурова Н.А. Об опыте создания фонетически аннотированного корпуса русской спонтанной речи // Фонетика сегодня: Материалы докладов и сообщений VI международной научной конференции 8-10 октября 2010 года.– М.: Институт русского языка им. В.В. Виноградова РАН, 2010.– С. 27–28.
- Венцов А.В., Слепокурова Н.А., Риехакайнен Е.И., Апушкина И.Е., Корешкова Е.И. Из опыта работы с русской спонтанной речью: создание фонетически транскрибированных текстов // X выездная школа-семинар "Проблемы порождения и восприятия речи": Материалы.– Череповец: ГОУ ВПО "Череповецкий государственный университет", 2011.– С. 169–179.
- Венцов А.В., Слепокурова Н.А. Фонетика русского звучащего текста и проблема моделирования процессов восприятия речи // Человек говорящий: исследования XXI века: коллективная монография / Под ред. Л.А. Вербицкой, Н.К. Ивановой.– Иваново: ФГБОУ ВПО "ИГХТУ", 2012.– С. 43–50.
- Кипяткова И.С. Обзор подходов к моделированию спонтанной речи // Труды второго междисциплинарного семинара "Анализ разговорной русской речи" (АРЗ – 2008).– СПб.: ГУАП, 2008.– С. 70–77.
- Кузнецов В.Б. О принципах акустической классификации русских гласных // Язык и речь: проблемы и решения: Сборник научн. трудов к юбилею проф. Л.В. Златоустовой / Под ред. Г.Е. Кедровой и В.В. Потапова.– М.: МАКСПресс, 2004.– С. 100–116.
- Лобанов Б.М., Цирульник Л.И. Моделирование внутрисловных и межсловных фонетико-акустических явлений полного и разговорного стилей речи в системе синтеза речи по тексту «Мультифон» // Труды первого междисциплинарного семинара "Анализ разговорной русской речи" (АРЗ – 2007).– СПб.: ГУАП, 2008.– С. 57–71.
- Bertrand R., Blache P., Espesser R., Ferre G., Meunier C., Priego-Valverde B., Rauzy S. Le CID - Corpus of Interactional Data-: protocoles, conventions, annotations // Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA).– 2006.– Vol. 25.– P. 25–55.

- Dilley L.C., Pitt M.A. A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition // J. Acoust. Soc. Am.– 2007.– Vol. 122, N 4.– P. 2340–2353.
- Gillis, S. 2001. Protocol voor brede fonetische transcriptie // http://tst-centrale.org/images/stories/producten/documentatie/cgn_website/doc_English/topics/project/phonetics/index.htm
- Kassevich V. B., Ventsov A. V., Yagounova E. V. The simulation of continuous text perceptual segmentation: A model for automatic segmentation of written text // Language and Language Behavior.– 2000.– V. 3, Pt. II.– SPb.– P. 48–59.
- Leitner C., Schickbichler M., Petrik S. Example-based Automatic Phonetic Transcription // Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Paris, European Language Resources Association (ELRA), May, 2010. (<http://www.spsc.tugraz.at/tools/extra-example-based-automatic-phonetic-transcription>)
- Schuppler, B. Automatic Analysis of Acoustic Reduction in Spontaneous Speech.– Nijmegen: Radboud University, 2011.
- Warner N. Methods for studying spontaneous speech // The Oxford Handbook of Laboratory Phonology / Ed. by A.C. Cohn, C. Fougerson, and M.K. Huffman.– 2012.– P. 621–633.