УДК 004.934

В.В. Робейко

МОДЕЛИРОВАНИЕ ОСОБЕННОСТЕЙ СПОНТАННОЙ УКРАИНСКОЙ РЕЧИ В СИСТЕМАХ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧЕВОГО СИГНАЛА

Исследованы особенности спонтанной украинской речи с учетом их дальнейшего моделирования в процессе распознавания речи. Отдельное внимание уделяется акустической, фонетической и лексической компонентам системы распознавания речевого сигнала, прогнозированию ударений в словах и учету признаков спонтанности во время формирования речевых и текстовых корпусов для обучения системы. Предложенная базовая экспериментальная система распознавания спонтанной речи в реальном времени оперирует словарем до ста тысяч слов и позволяет набирать текст под диктовку.

Введение. Дикторонезависимое распознавание слитной спонтанной речи в реальном времени позволяет решать широкий спектр прикладных задач в самых разных областях человеческой жизни. Анализ патентов коммерческих фирм и публикаций известных научных центров мира показывает, что в последнее время появилось много программных средств диктования на ПК, а также сетевые сервисы, позволяющие устно формировать поисковые запросы или диктовать письма электронной почты. Изолированно произнесенные слова и слитная подготовленная речь (например, чтение новостей дикторами телевидения) в настоящее время может распознаваться с пословной точностью около 95 % [1]. Результаты распознавания спонтанной речи в реальных условиях общения значительно хуже. Построение качественно работающей системы распознавания для спонтанной речи требует не только соответствующего материала для обучения, но и учета особенностей такой речи при разработке акустических и лингвистических компонентов систем распознавания.

Интерес к спонтанной речи как к основному типу коммуникации между людьми появился в научном мире сравнительно недавно, а в сфере информационных речевых технологий — не более двух десятилетий назад [2]. Даже сейчас наши знания о структуре спонтанной речи не настолько соответствуют действительности, чтобы достичь необходимого прорыва в ее распознавании. Поэтому в работах по распознаванию речевого сигнала чрезвычайно актуальны исследования и построение моделей, которые будут учитывать свойства спонтанной речи. Развитие инструментальных средств обработки речи и текста открывает дорогу к созданию систем распознавания речи, все более отвечающих растущим требованиям пользователя.

Цель исследования — поиск мер по повышению точности распознавания речи с учетом спонтанного характера речи. Необходимо описать выявленные особенности спонтанной украинской речи, определить наиболее важные из них с точки зрения влияния на точность распознавания речи и предложить решения проблем, связанных с моделированием особенностей спонтанной речи.

Характеристики спонтанной речи. Под спонтанной речью как устной, так и письменной, понимается речь неподготовленная (или подготовленная минимально), осуществляемая говорящим в постоянно меняющихся коммуникативных условиях [3]. Спонтанная речь преобладает в реальной

речевой коммуникации, именно поэтому она является доминирующей, основной, первичной по сравнению с подготовленной речью (чтение, декламация). В неподготовленности и повышенной ситуативной обусловленности спонтанного типа речи состоит основная сложность его анализа. При спонтанном порождении речи ее организация на уровне фонетических, лексических и синтаксических единиц отличается от изученных и описанных исследователями явлений.

Спонтанная речь труднее поддается автоматическому распознаванию в первую очередь из-за своей вариативности, которая проявляется на аллофонном и фонемном уровне. В потоке слитной речи одни и те же слова произносятся по-разному: меняется место ударения, появляется редукция, недоговариваются окончания или накладываются конец одного слова и начало другого. Спонтанная речь является менее стабильной по своим характеристикам по сравнению с чтением, что проявляется как в качественных, так и в количественных особенностях гласных, например в меньшем различии между ударными и безударными гласными. Причинами такого явления можно назвать менее выраженные интонационные центры, нарушенную акцентно-ритмическую структуру синтагм, наличие внутрисинтагменных пауз хезитации, уменьшенное количество фонетических слов в синтагме и др.

Особенности спонтанной речи проявляются также в нарушениях плавности речи, выраженных в виде фальстартов, самоисправлений говорящего, сбоев в употреблении грамматических форм и синтаксического порядка слов, множестве недоговоренных слов и фраз, повторах, частой смене темпа речи и т.д. Исследования подтверждают, что вышеперечисленные признаки значительно влияют на точность распознавания речи [4]. Не менее важен характер лексики спонтанной речи (использование диалектов, суржика, грубой и ненормативной лексики).

Спонтанная речь практически на всех языковых уровнях в значительной степени обусловлена социальными характеристиками говорящих. Еще одной важной особенностью спонтанной речи можно считать непостоянность ее признаков, которые проявляются в разной степени в зависимости от индивидуальности диктора. Например, одним дикторам мало свойственны паузы хезитации, другие — активно используют колебания и повторы, третьи — редуцированное или «растянутое» произношение. К тому же речь дикторов часто нарушена такими экстралингвистическими неинформативными явлениями как шумное дыхание, кашель, смех, причмокивание и т.д.

Отличительной чертой спонтанной речи украинцев является многоязычность устных текстов и наличие суржика (наличие в речи элементов нескольких языков). Даже в речи одного диктора часто перемешиваются украинский и русский языки.

Письменная спонтанная речь имеет свои особенности: нестандартная или неоднозначная транслитерация текстов, неправильная орфография, нестандартные сокращения и аббревиатуры, специфическое построение предложений, ненормативная лексика, суржик, наличие эмотиконов (смайликов) [5]. Примеры письменной спонтанной речи можно встретить, анализируя СМС-сообщения, чаты, форумы, блоги, личную электронную переписку.

Оценивание параметров генеративной модели распознавания спонтанной речи. Входящий сигнал спонтанной речи преобразуется в последовательность акустических векторов-признаков $\boldsymbol{Y}_{1:T} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_T)$ в результате препроцессинга. Затем декодер пытается отыскать последовательность речевых сегментов, заданных языковыми символами, $\mathbf{w}_{1:L} = (w_1, w_2, ..., w_L)$, которая наиболее вероятно соответствует наблюдаемому $\boldsymbol{Y}$:

$$\tilde{\tilde{\boldsymbol{w}}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, P(\boldsymbol{w} \mid \boldsymbol{Y}) \cong \underset{\boldsymbol{w}}{\operatorname{argmax}}\, p(\boldsymbol{Y} \mid \boldsymbol{w}) P(\boldsymbol{w}).$$

Эквивалентность правой части выражения, вытекающая из применения правила Байеса, представляет базовую формулировку генеративной модели распознавания речи. Акустическая — $p(\boldsymbol{Y} \mid \boldsymbol{w})$ — и лингвистическая — $P(\mathbf{w})$ — составляющие генеративной модели описываются каждая своими стохастическими порождающими грамматиками.

Акустическая модель формируется в результате композиции моделей базовых речевых элементов q. Для моделирования экстралингвистических явлений, свойственных спонтанной речи, в алфавит базовых элементов дополнительно к фонемам и фонемам-паузам вводятся символы, отображающие неинформативные звуки.

Композиция базовых элементов управляется словарем произношений $\boldsymbol{V}$ речевых сегментов w, которыми являются информативные и неинформативные слова.

Свойственные спонтанной речи отклонения от нормы произношения частично могут моделироваться на акустическом уровне. На рис. 1 изображена стохастическая порождающая грамматика для слова «нього», полученная в результате композиции акустических моделей четырех фонем (*н'*, *о*, *г*, *о*). Цифры рядом со стрелочками означают количество отсчетов времени, за которое производится переход в следующее состояние, исходя из того, что крайние состояния фонем являются неэмитентными, т.е. переход из них осуществляется за нулевой промежуток времени. Сумма вероятностей осуществления перехода из любого состояния равна единице.

Заметим, что фонема *г* в контекстах, аналогичных слову «нього», зачастую реализуется настолько слабо, что практически выпадает. Поэтому целесообразно ввести в топологию акустической модели соответствующей контекстнозависимой фонемы — фонемы-трифона *ого* — допустимые переходы между состояниями, позволяющие сократить до минимума продолжительность фонемы. На рисунке дополнительно добавлены состояния, одно из которых позволяет сократить длительность фонемы до двух отсчетов, а другое и вовсе допускает выпадение фонемы из потока речи.

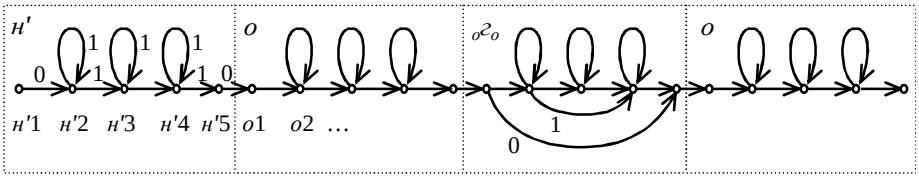

Рис. 1. Композитная модель слова, состоящего из четырех фонем

Особенности **лингвистической модели** для спонтанной речи состоят в необходимости обработки заполненных пауз, недоговоренных слов и слов, произнесенных по частям, например посложно. Предположим, что

заполненные паузы и обрывы слов отнесены к словарю «прозрачных» слов $V_{\text{прозр.}}$, частичные слова находятся в словаре $V_{\text{част.}}$, все остальные слова — в словаре V. Тогда при поступлении слова из $V_{\text{прозр.}}$ или $V_{\text{част.}}$ накопление гипотезы биграммы слов откладывается в первом случае до поступления следующего слова из V, а во втором — до окончания накопления полного слова.

На рис. 2 изображен граф, по которому формируется биграмма $P(w_j \mid w_i)$ для ее применения в лингвистической составляющей декодера. Точечные стрелки указывают на поступление «прозрачного» слова, пунктирные — частичного слова. Символ **0** означает конкатенацию сегментов слова. В общем случае возможно движение одновременно по двум стрелкам, когда полное слово может совпадать с частичным: $w \in V \cap V_{\text{част.}} \neq \varnothing$.

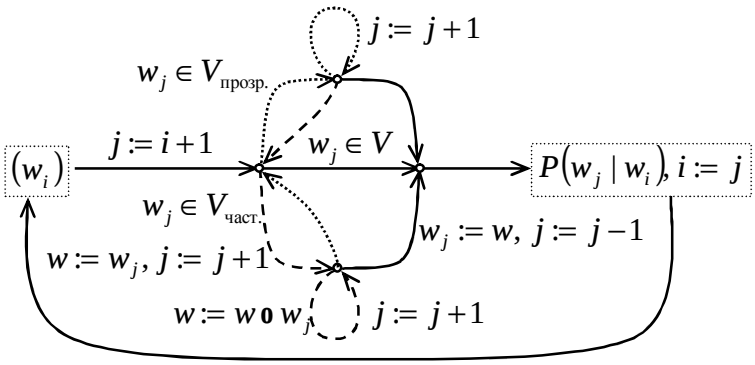

Рис. 2. Формирование аргумента для биграммы лингвистической модели с учетом «прозрачных», частичных и посложно произнесенных слов

На практике словарь частичных слов очень громоздок и имеет смысл ограничиться лишь наиболее частотными словами.

Словарь распознавания, учитывающий спонтанное произнесение. Графемно-фонемные преобразования нужны для формирования словарей произношения при оценке параметров акустической модели. Именно в этих словарях должна быть отражена вариативность произношения на фонемном уровне, свойственная спонтанной речи [6, 7]. В ходе работы по распознаванию речевого сигнала на основе анализа спонтанной речи нескольких сотен дикторов была разработана система, на вход которой подается орфографический текст, содержащий только символы из алфавита букв, включая символы границы между словами и морфемами и обозначениями ударения. На выходе системы получаем последовательности фонем, соответствующие различным вариантам произношения входящего текста [8]. Разработанная система многозначного транскрибирования орфографических текстов использует конечный автомат, который предусматривает возможность таблично задавать контекстно зависимые правила преобразований одних обобщенных последовательностей символов в другие. При этом в каждом правиле задается ширина шага, по которому происходит переход к следующей последовательности символов. Для

построения транскрипции украиноязычных текстов достаточно не более 30–35 правил. Применение многих правил позволяет генерировать сразу несколько вариантов транскрипции одного и того же слова или генерировать нужный вариант из нескольких возможных, например описывая спонтанную речь говорящего или группы дикторов.

Возможность генерировать сразу несколько вариантов транскрипции одного и того же слова позволяет продемонстрировать в словаре вариативность произношения наиболее частотных украинских слов, редуцирование и растяжение слов во время быстрого темпа речи, нечеткое произношение и подобные явления наряду с литературным вариантом произношения. Также система транскрибирования позволяет генерировать транскрипции для таких специфических подсловарей, как словарь суржика, социальных и территориальных диалектов, аббревиатур и др. Введение нескольких способов произнесения слов в словаре в целом улучшает надежность распознавания спонтанной речи [9]. Пример многовариантной транскрипции можно увидеть в табл. 1.

Таблица 1.

Пример многовариантной транскрипции слова «абсолютно»

Слово	Транскрипция	Пояснения к транскрипции, применение разных правил
абсолютно	а б с о л' У т н о	литературный вариант транскрипции
	а б с о л' у т н о	безударный вариант
	а б с а л' У т н а	редукция гласного «о»
	а п с о л' У т н о	ассимиляция по глухости
	а п с а л' У т н а	редукция гласного «о» + ассимиляция по глухости
	а п с а л' у т н а	безударный вариант + редукция гласного «о» + ассимиляция по глухости

Прогнозирование ударений в словах. Спонтанная речь характеризируется повышенной динамичностью на лексическом уровне. Постоянно появляются новые слова и выражения, интенсивно используется диалектная и суржиковая лексика, ненормативная лексика. Существующие и используемые словари ударений [10] не в состоянии зафиксировать и передать многообразие лексических форм, а задействование экспертов для составления дополнительных словарей ударений связано со значительными трудовыми затратами. Предлагается использовать алгоритм, в котором решение о месте ударения в слове принимается на основе знаний об ударениях в оговоренном словаре ударений и с использованием массива текстов [11].

Фиксируется некая сегментация S, задающая разбиение заданного слова на сегменты S_i, которые учитывают возможное ударение. Каждой такой сегментации сопоставляется критерий, вытекающий из встречаемости сегментов в текстовом корпусе. Методом динамического программирования осуществляется направленный перебор допустимых сегментаций с целью нахождения одной или более сегментаций с лучшим критерием, по которым восстанавливаются позиции ударений.

На рис. 3 проиллюстрировано действие алгоритма на примере слова «обама», отсутствующего в базовом словаре украинского языка. Слово

представляется в виде конкатенации допустимых сегментов S_i, содержащих гипотетическую ударную гласную (заглавная буква), длиной не более четырех символов (включая символы начала и конца слова). Вторая сверху строка содержит поступающий символ, порождающий множество перечисленных под этим символом допустимых сегментов, которые дополнены значением частичного критерия. Плавные стрелки берут начало в предшествующих потенциально оптимальных сегментах. Для сегментов, к которым стрелки не проведены, указывается имя оптимального предыдущего сегмента. Сегменты и стрелки, отвечающие оптимальному разбиению, выделены утолщением.

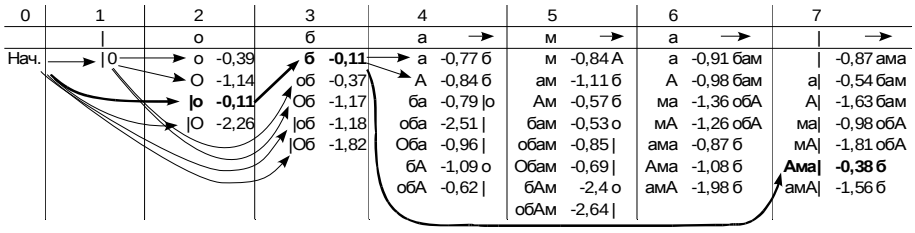

Рис. 3. Предсказание ударения для слова «обама», не вошедшего в базовый словарь.

Существуют трудности при объективной оценке надежности предсказания ударений. Например, слово может иметь допустимое ударение в различных позициях. Тем не менее эксперименты показывают ошибку предсказания ударений на уровне 5–10 % относительно количества слов, не вошедших в базовый словарь.

Речевой и текстовый материал для исследований. Любая система распознавания речи требует предварительного этапа обучения или настройки. Для этого необходимы речевые и текстовые материалы, близкие к условиям реальной эксплуатации. Акустические модели фонем составляются путем предварительного обучения системы на большом массиве данных (речевом корпусе), который должен включать от нескольких десятков до сотен часов устной речи, записанной от многих сотен дикторов [2, 12]. Важно создать такой речевой корпус, который будет отображать современную языковую ситуацию в Украине и соответствовать специфическим задачам распознавания спонтанной речи. Акустический корпус украинской эфирной речи (АКУЭР) [13] — общий по цели своего применения акустический корпус, который содержит подготовленную и спонтанную речь (последняя составляет бóльшую часть корпуса). Все материалы корпуса по типу речевого сигнала разделяются на теле- и радиовещание, также присутствуют небольшие вкрапления записи публичной речи и речи в естественной среде. Основные языки материалов корпуса — украинский и русский. В АКУЭР вошли материалы разной тематики и жанров, но основу корпуса составили звуковые записи рубрик: новости и интервью (политика, культура, образование, общество и т.д.), телепередачи и телетрансляции (судебные заседания, политические дебаты, публичные выступления и др.). Среди записей встречается речь дикторов разного возраста, пола, социального положения и профессий, что отражает состав дикторов телевизионного эфира. АКУЭР характеризуется следующими количественными

показателями: более 300 часов аннотированной речи, произнесенной более чем 2 000 дикторов. Словарь корпуса содержит более 65 000 слов украинского языка и почти 60 000 слов русского языка.

Весь материал корпуса был размечен (аннотирован) экспертами таким образом, чтобы отобразить разделение слитного потока речи на небольшие отрывки, удобные для входа в систему обучения распознаванию. Также эксперты указали разнообразные фоновые (наложенные на речь) и изолированные лингвистические и экстралингвистические явления в речевых сегментах (описание языка, способа произношения слов, шума, неинформативных слов и звуков, которые произносит диктор, диалектной и суржиковой лексики и т.п.).

Лингвистическая модель обучается путем предварительного анализа большого массива текстов и построения вероятностей следования одних слов за другими. Корпус текстового материала был создан на основе скачанных из Интернета украиноязычных текстов разных тематик и жанров (новости и публицистика, художественные тексты, энциклопедические статьи, юридические тексты и др.) [14]. Для моделирования особенностей спонтанной речи в текстовый корпус были включены материалы разделов обсуждений и комментариев пользователей на новостных сайтах. Общий объем текстового корпуса — 2 Гб текстов (250 млн реализаций слов), пропущенных через текстовый фильтр (числа и символы были преобразованы в слова, удалены всевозможные повторения и т. п.).

Экспериментальная система распознавания спонтанной украинской речи в реальном времени. Для создания базовой системы преобразования речевого сигнала в текст использовались как собственные разработки, так и разнообразные программные инструментарии, доступные в Интернете: ***HTK***, ***Julius***, ***MITLM*** и др. На рис. 4 изображена общая структура автоматического распознавателя спонтанной речи.

В *модуль реального времени* поступает *речевой сигнал* через один из доступных источников (микрофон или файл). При прохождении через *детектор голосовой активности* сигнал разбивается на сегменты по признаку наличия голосового ввода. Используются простые акустические признаки в амплитудно-временном пространстве на основе текущей амплитуды и количества переходов через ноль. *Блок препроцессора* переводит сигнал в пространство первичных векторов-признаков. При этом используется мел-кепстральное преобразование с вычитанием среднего значения. *Декодер* производит сравнивание входящего речевого сегмента с гипотезами эталонного сигнала допустимых последовательностей словарных сегментов из *рабочих словарей*, применяя некую осторожную стратегию отбрасывания малоперспективных гипотез. Для этого используются данные из *акустической* (40 часов аннотированной речи из АКУЭР) и *лингвистической* (88,5 млн триграмм для рабочего словаря на 100 тыс. слов) моделей. Последовательность слов, по которой генерируется наиболее похожий эталонный сигнал, объявляется *ответом распознавания*.

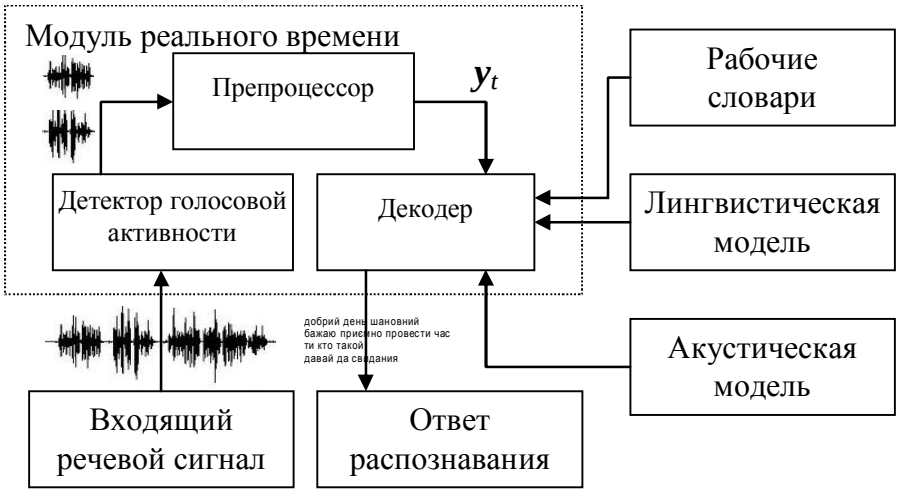

Рис. 4. Структура автоматического распознавателя спонтанной речи

Базовая система преобразования речевого сигнала в текст используется для экспериментальных исследований распознавания спонтанной речи, а добавленный графический интерфейс пользователя позволяет демонстрировать распознавание спонтанной речи в реальном времени на ПК и проводить опытную эксплуатацию системы [14, 15].

Условия эксплуатации разработанной системы учитывают ожидания потенциального пользователя. Словарь системы охватывает общепринятую лексику и множество слов некоторых предметных областей: например, естественные науки, строительство, медицина, юриспруденция и т.д. В нашем случае выбрана новостная тематика (политика, экономика, культура, спорт и погода). На акустическом уровне, система воспринимает речь любого адекватного пользователя. Заранее подготовленная речь, прочитанные тексты, спонтанные высказывания распознаются на одном уровне. Сильно зашумленные записи и перекрытия речи разных лиц в одном канале записи на данном этапе разработок не распознаются с допустимой для эксплуатации точностью.

Во время опытной эксплуатации этой системы использовались словари на 10, 20, 50 и 100 тысяч слов. Поскольку для всех словарей декодирование происходило в реальном времени (до 15 % на процессоре *i*7), было проведено более детальное исследование максимального словаря в 100 тысяч слов. Система тестировалась в качестве машины набора текста под диктовку пятнадцатью экспертами (в режиме чтения текстов и спонтанного монолога). В условиях эксплуатации, описанных выше, пословная ошибка распознавания составляет в среднем 10 % для обоих типов речи. Для спонтанной речи такой показатель надежности распознавания достигнут за счет реализации большинства предложенных моделей.

Направления будущих исследований включают: увеличение объема словаря распознавания до одного миллиона слов с целью покрытия практически всего лексикона произвольной речи, оптимизацию лингвистической модели путем введения классов слов, использование контекстнозависимых моделей фонем, кластеризацию дикторов и

настраивание на голос диктора, прогнозирование знаков пунктуации и регистра слов с последующей смысловой реконструкцией распознаваемого текста.

Выводы. Предложенные в работе математические модели анализа речевого сигнала позволяют учитывать свойства спонтанной речи на акустическом, фонетическом и лексическом уровнях. Применение как разработанных в ходе исследований, так и общедоступных инструментальных средств с использованием речевых и текстовых баз данных и знаний привело к созданию базовой системы распознавания речевого сигнала, ориентированной на спонтанную речь. Проведенные эксперименты показали прикладную ценность исследований, которая выражается в увеличении надежности и скорости распознавания. Опытная эксплуатация системы диктовки текстов демонстрирует возросшие потребительские качества разработанной речевой технологии, востребованность и перспективность дальнейших разработок в области распознавания и синтеза спонтанной речи.

1. *Gales* M., *Young S.* The Application of Hidden Markov Models in Speech Recognition // Foundations and Trends in Signal Proc. — 2007. — **1**(3). — P. 195–304.
2. *Furui S., Nakamura M., Ichiba T., Iwano K.* Why is the Recognition of Spontaneous Speech so Hard? // Text, Speech and Dialogue, ser. Lecture Notes in Artificial Intelligence. — 2005. — P. 9–22.
3. *Бондарко Л.В.* Спонтанная речь и организация системы языка // Бюллетень фонетического фонда русского языка. — Санкт-Петербург, 2001. — № 8. — С. 17–23.
4. A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models / M. Adda-Decker, B. Habert, C. Barras et al. // Proc. of Disfluency in Spontaneous Speech Workshop, Göteborg, Sweden. — 2003. — P. 67–70.
5. *Людовик Т.В., Робейко В.В.* Озвучивание SMS-сообщений, отправляемых на стационарные телефоны // Речевые технологии. — 2009. — Вып. 3 — С. 24–33.
6. *Кипяткова И.С.* Обзор подходов к моделированию спонтанной речи // Труды Второго междисциплинарного семинара «Анализ разговорной русской речи» (АРЗ-2008). — Санкт-Петербург, 2008. — С. 70–77.
7. *Amdal I.* Learning pronunciation variation. A data-driven approach to rule-based lexicon adaptation for automatic speech recognition. PhD thesis. — Department of Telecommunications Norwegian University of Science and Technology. Norway. — 2002. — 182 p.
8. *Робейко В.В., Сажок М.М.* Багатозначна багаторівнева модель перетворення орфографічного тексту на фонемний // Штучний інтелект. — 2011. — № 4. — С. 117–125.
9. *Людовик Т.В., Робейко В.В., Пилипенко В.В.* Автоматическое распознавание спонтанной украинской речи (на материале акустического корпуса украинской эфирной речи) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной междунар. конф. «Диалог» (Бекасово, 25–29 мая 2011 г.). — Москва, 2011. — Вып. 10 (17). — С. 478–488.
10. *Широков В.А., Манако В.В.* Організація ресурсів національної словникової бази // Мовознавство.— 2001. — № 5. — С. 3–13.
11. *Робейко В.В., Сажок М.М.* Використання текстового корпусу для прогнозування наголосів у словах української мови // Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту: Матеріали міжн. наук. конф. — Херсон, 2012. — С. 171–172.
12. *Пилипенко В.В., Робейко В.В.* Автоматизированный стенограф украинской речи // Искусственный интеллект. — 2008. — № 4. — С. 768–775.

13. Створення акустичного корпусу українського ефірного мовлення / Н.Б.Васильєва, В.В. Пилипенко, О.М. Радуцькийи та ін. // Обробка сигналів і зображень та розпізнавання образів: Десята Всеукраїнська міжнар. конф. — Київ, 2010. — С. 55–58.
14. *Робейко В.В., Сажок М.М.* Розпізнавання спонтанного мовлення на основі акустичних композитних моделей слів у реальному часі // Штучний інтелект.— 2012. — № 4. — С. 253–263.
15. www.cybermova.com/products/stt-demo.htm.

Международный научно-учебный центр
информационных технологий и систем
НАН Украины и Министерства образования
и науки, молодежи и спорта Украины, Киев

Получено 27.11.2012