УДК 004.394

О.Н. Ладошко, А.Н. Продеус, канд. техн. наук

Разметка спонтанной украинской речи

Представлена классификационная схема особенностей спонтанной украинской речи. Разработана расширенная система разметки таких особенностей. Представлены технология и алгоритмы автоматизированного поиска нарушений спонтанной речи.

The classification scheme of particularities Ukrainian spontaneous speech features is represented. Extended annotation scheme of the features was developed. Technology and algorithms of automated search of spontaneous speech disfluencies are represented.

Ключевые слова: *разметка, особенности спонтанной речи, создание речевого корпуса.*

Введение

Изучению и моделированию особенностей, отличающих подготовленную устную речь от неподготовленной, а также речевым сбоям в спонтанной речи посвящено значительное количество трудов международных конференций и семинаров [1, 2, 3]. В странах СНГ исследованием нарушений спонтанной речи (речевых сбоев) занимались преимущественно лингвисты [4, 5, 6], а детальному изучению структуры элементов спонтанной русской речи для их моделирования в системах автоматического распознавания и синтеза речи по тексту посвящено лишь несколько работ [7, 8].

Украинскими исследователями в последнее время ведутся работы по созданию системы автоматического распознавания спонтанной украинской речи (АРСУР) [9]. Заметим, что под *спонтанной речью* понимается *разговорная речь* во всех её степенях подготовленности. Экспериментальные исследования свидетельствуют, что надёжность распознавания спонтанной речи, с учетом зависимости частот встречаемости слов и использования биграммной модели языка, даёт незначительное увеличение надёжности распознавания при увеличении размера словаря (для семи частотных словарей из работы [9] надёжность повышается в среднем на 1,07%). Для улучшения надёжности распознавания в работе [9] предложено вводить индивидуальные модифицированные транскрипции. Подобные транскрипции, рассмотренные в [7] и именуемые *альтернативными транскрипциями,* предназначены для устранения несоответствия между наблюдаемым произношением и принятыми фонетическими транскрипциями.

Однако введение модифицированных транскрипций для выделенных групп диктор- ов [9] также привело к незначительному повышению надежности распознавания (в среднем на 0,92% по выборке из таблицы 7, 0,78% для всех выборок работы [9] и на 0,47% отдельно для каждого диктора), а в некоторых случаях было обнаружено, что надёжность АРСУР даже может ухудшиться.

К сожалению, для накопленных в библиотеках звуковых и текстовых материалов украинской речи [9] до настоящего момента была реализована лишь разметка звуковых фрагментов в соответствии с текстовым сопровождением, что не позволяет изучать и моделировать элементы спонтанной украинской речи. По нашему мнению, решению проблемы повышения качества системы АРСУР [9] может существенно помочь введение специальной системы разметки, которая позволила бы отделить «чистую» речь от особенностей спонтанной речи [10, 11].

Цель данной работы состоит в дальнейшем развитии системы разметки спонтанной украинской речи, предложенной в [10,11].

1. Создание речевого корпуса

Работа по созданию аннотированного корпуса “живой” украинской речи проводилась авторами с конца 2009 года. В дальнейших исследованиях были выявлены и статистически исследованы особенности спонтанной речи дикторов и влияние основной особенности спонтанной речи – *речевых сбоев* [10, 11] – на АРСУР. Эти исследования позволили повысить надежность АРСУР на 1…8%, в зависимости от детализации производимой разметки [10]. В работах [10, 11], были представлены только основные определения выявляемых нарушений речи (речевых сбоев) и их примеры.

2. Методика подготовки речевого корпуса

2.1 Анализируемые материалы и программное обеспечение

Для исследований использовались стенограммы заседаний Верховной Рады Украины. Общая длительность анализируемых звукозаписей составила около 20 часов, количество

различных дикторов - 203. В таблице 1 приведены оценки количественных характеристик анализируемых материалов.

Заметим, что анализируемым материалам свойственны такие особенности как темповая неоднородность, усиление редукции (частое изменение звуков, состоящее в утрате полноты их звучания), достаточно высокое качество записи (каждое место оснащено микрофоном).

Тексты стенограмм были представлены в виде предварительно сегментированного на фразы текста в соответствии со звуковым сопровождением [9].

Автоматизация аннотирования *текста стенограммы* осуществлялась с помощью специальной программы тестирования [9], графический интерфейс которой показан на рис. 1.

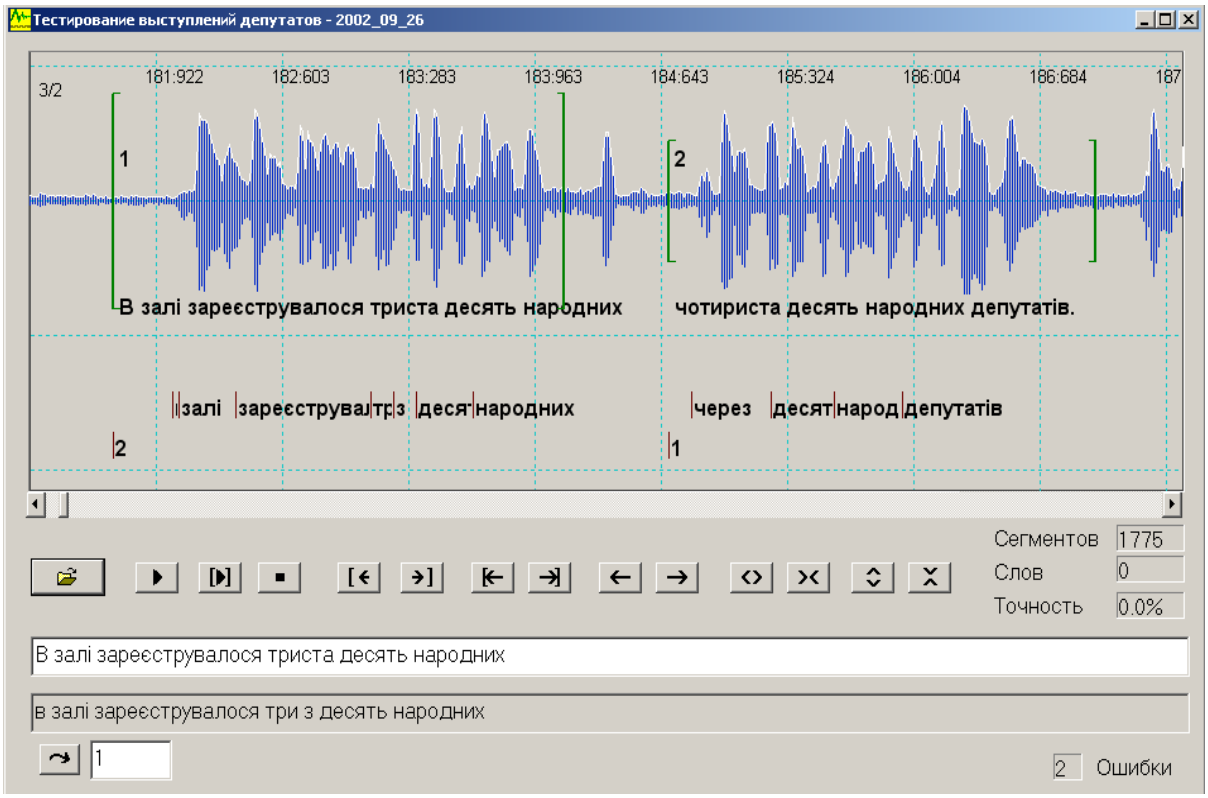

Рис. 1. Графический интерфейс программы тестирования украинской речи

Таблица 1. Характеристики исследуемого материала

№ п/п	Имя файла	Кол-во сегмен-тов, шт.	Кол-во слов, шт.	Кол-во сегментов с нарушения-ми, шт.	Общее кол. дикторов, число	Общее время, с	Общее кол. наруше-ний, шт.
1	2002_09_26	1775	8223	490	52	3809	494
2	2002_10_08	3087	14109	743	43	6619	758
3	2002_10_16	3640	15343	812	47	6899	824
4	2002_10_18	3188	13219	562	43	5873	569
5	2002_10_22	6129	20641	893	57	9947	906
6	2002_10_23	3494	14482	606	45	6631	608
7	2002_10_24	2889	15512	926	54	6686	944
8	2002_10_25	4575	15967	794	53	7480	806
9	2002_11_12	3624	13750	651	46	6797	672
10	2002_11_19	2982	11786	472	42	5355	473
11	2002_11_20	3623	13506	619	40	6268	622
Сум.		39006	156538	7568	522	72364	7676

2.2 Аннотирование речевого корпуса

В ходе обработки корпусов спонтанной речи мы неизбежно сталкиваемся со "свободой" и вариативностью произношения каждого человека [4, 7, 10]. Именно вольность формирования спонтанной речи является первопричиной необходимости введения этапа предварительной подготовки «сырых» речевых данных для системы АРСУР.

В таблице 2 приведен перечень нарушений речи, выявляемых на этапе предварительной обработки спонтанной речи. Информация, о частотном составе всех выявленных нарушений речи, представлена в последнем столбце этой таблицы.

Прежде всего, выявлялись *звуки, производимые самим диктором* (дыхание, кашель), которые выполняют вспомогательную функцию *пауз* в процессе речеобразования. Другими выявляемыми *паузами*, влияющими на вариативность речевого сигнала, были: *заполненные* или *вокализованные паузы* (воспроизводимые с участием голоса диктора); *растягивания звуков* (не по правилам их произношения под ударением), выполняющие функцию своеобразных заполнителей промежутков речи.

Кроме того, негативное влияние на систему АРСУР оказывают: *незаконченные слова* (редуцированные и нередуцированные обрывы без самоисправлений, фальстарты, различного рода коррекции); *непреднамеренные повторы слов*, повторы со вставкой; различной степени *редукции словоформ* [10, 11].

Кроме вышеперечисленных явлений, в речи политиков наблюдается одновременное употребление слов на нескольких языках - так называемый *суржик*. Сложность распознавания суржика заключается в отсутствии норм и правил, согласно которым слова можно отнести к украинскому или русскому языку.

Исследования также свидетельствуют о сложности распознавания *аббревиатур*. Для устранения этой проблемы необходимо создавать и постоянно пополнять дополнительные словари всех *слов-аббревиатур.*

К нарушениям речи диктора отнесена также речь людей, находящихся недалеко от микрофона диктора, а также все фрагменты речи, записанные с малым уровнем.

Еще одной проблемой является вычленение понятия *предложения* в тексте спонтанной речи. Замечено, что использование подготовленного речевого материала с неверно определенными границами сегментов может привести к ухудшению работы системы АРСУР. Аналогичные особенности были обнаружены и в начальных позициях исследуемых сегментов. Следует учесть, что степень неоднозначности сегментирования речевого потока может возрасти по мере возрастания степени неофициальности общения.

Как показали исследования, присутствие в спонтанной речи *речевых сбоев* может стать дополнительным источником информации для решения проблемы сегментации речевого потока. К таким сбоям относятся: *артефакты речи* (например, цоканье и причмокивание языком); *дискурсивные слова* (например, рус. язык: «ну», «ага», «угу», «да», «но», «что»; укр. язык: «що», «коли», «ну», «але»). Поскольку речевые сбои неразрывно связаны с прерыванием речевого потока либо с изменением его акустических характеристик, можно предположить, что задача предварительной сегментации речевого сигнала сводиться к задаче обнаружения речевых сбоев. Например, в ряде случаев решение задачи сегментации речевого потока сводиться к построению детектора явлений, а именно *детектора речевых сбоев*.

Таблица 2. Перечень особенностей и система их аннотации

№	Особенность	Символьное обозначение (шаблон)	Примеры явлений в текстах стенограмм	Кол-во, шт.
1	заполненные паузы	(а), (е), (є), (у), (о), (м), (н), (и), (і), (я)	[передбачені (е) вами] [окремих суб'єктів (а) цих операцій]	1077
2	растягивания звуков, откашливания, придыхания	(хе), (іе), (ее), (оа), (мм),(кг), (кх), (ау), (ае), (ен) (ме)	[Прошу (оа)] [(хе) то можна (хе)] [таким же чином (ме)]	433
3	слабая редукция	(слабо редуцированное слово)	[(вучені) Президенту України] [умовно (качи), Адміністрації]	1093

№	Особенность	Символьное обозначение (шаблон)	Примеры явлений в текстах стенограмм	Кол-во, шт.
4	лишние звуки в конце сегмента	[слово, завершающее сегмент (не)]	[Шановні коллеги (не)] [немає заперечень щодо…]	286
5	лишние звуки в начале сегмента	[(ть) лишний звук попадающий в сегмент]	[Ви знаєте, що цей паспорт видається на десять] [(ть) років і може бути продовжений…]	47
6	фальстарты или обмолвки	(Шан) Шановні фрагмент следующего слова	[(на) народному колезі] [(Шан) Шановні коллеги,] [корисно (об) обопільно]	126
7	повторы	явно не обозначались	[Прошу уваги, Прошу уваги]	111
8	повторы всех слов и групп слов в сегменте	явно не обозначались	[двадцять один двадцять один двадцять два тридцять п'ять]	480
9	редуцированные и нередуцированные обрывы без самоисправлений	слышимая часть слова (уточнит...@)	[деякі питання, які треба (уточнит...@)] [Є (потре…@) ? Будь ласка.]	163
10.	приглушенная речь с низким уровнем и разговор на фоне,	~(фрагмент тихой речи)	[я звертаю вашу увагу, ~(да да) що сьогодні…]	107
11	коррекция со вставкой (заполн. пауза или лексический материал)	&(обрыв) (вставка) самоисправление	[Для того, щоб (ее) &(більш), (е) меншість диктувала волю більшості?]	115
12	онлайн коррекция	^(обрыв) самоисправление	[Тому я ^(споча) почергово поставлю дві]	454
13	повтор со вставкой (заполн. пауза или лексический материал)	$(1-ый повтор) (вставка) 2-ой повтор	[$(слідчу комісію), створити тимчасову слідчу комісію.]	44
14	не вошедшие в сегмент повторы	%(первое повторяемое)	[Дякую, Петро (Миколайч). %(Шановні коллеги)]	36
15	мусор – посторонние звуки, в том числе из соседних сегментов	*(лишний элемент)	[шановні народні депутати *(з)] [з фракції Наша Україна, а я хочу нагадати:]	1296
16	сильно редуцированные слова или словосочетания	#(сильно редуцируемое слово)	[Тому просив би звернуть на це увагу і #(ппоувати) #(сконтилізувати) Верховну Раду.]	159
17	аббревиатуры	!(аббревиатура)	[Української !(РСР),]	289
18	суржик	{суржик}	[робочі місця, в кого {єсь}...]	798

2.3 Технология аннотирования

Аннотация явлений спонтанной речи состоит в такой разметке звуковых файлов, которая адекватно отображала бы нарушения речи (см. табл. 2), возникающие в реальных условиях речевой коммуникации.

Первичная аннотация явлений производилась вручную одним человеком-экспертом, что позволило сохранить единообразную организацию и стандартный способ представления в них данных. При этом в стенограммы проставлялись символы букв, которые отображали схожие со звучащими явлениями фонемы. Элементы, оказывающие влияние на систему автоматического распознавания речи [9] выделялись круглыми скобками (табл. 2). На данном этапе не осуществлялось разделение явлений на виды (рис. 3 – исходный пункт ветвления классификационного дерева).

На следующем, автоматизированном, этапе обработки спонтанной речи было предложено выделять (классифицировать) нарушения речи с применением специальных правил. Такие правила классификации нарушений речи отображены на рис. 3 в виде дерева утверждений и фактов наличия нарушений речи. Спускаясь вниз по ветвям классификационного дерева, можно проследить, по каким признакам производилась классификация нарушений спонтанной речи. Автоматизация обнаружения и классификации явлений спонтанной речи не только заметно разгрузила человека-эксперта от рутинной работы, но и позволила существенно сократить время аннотирования.

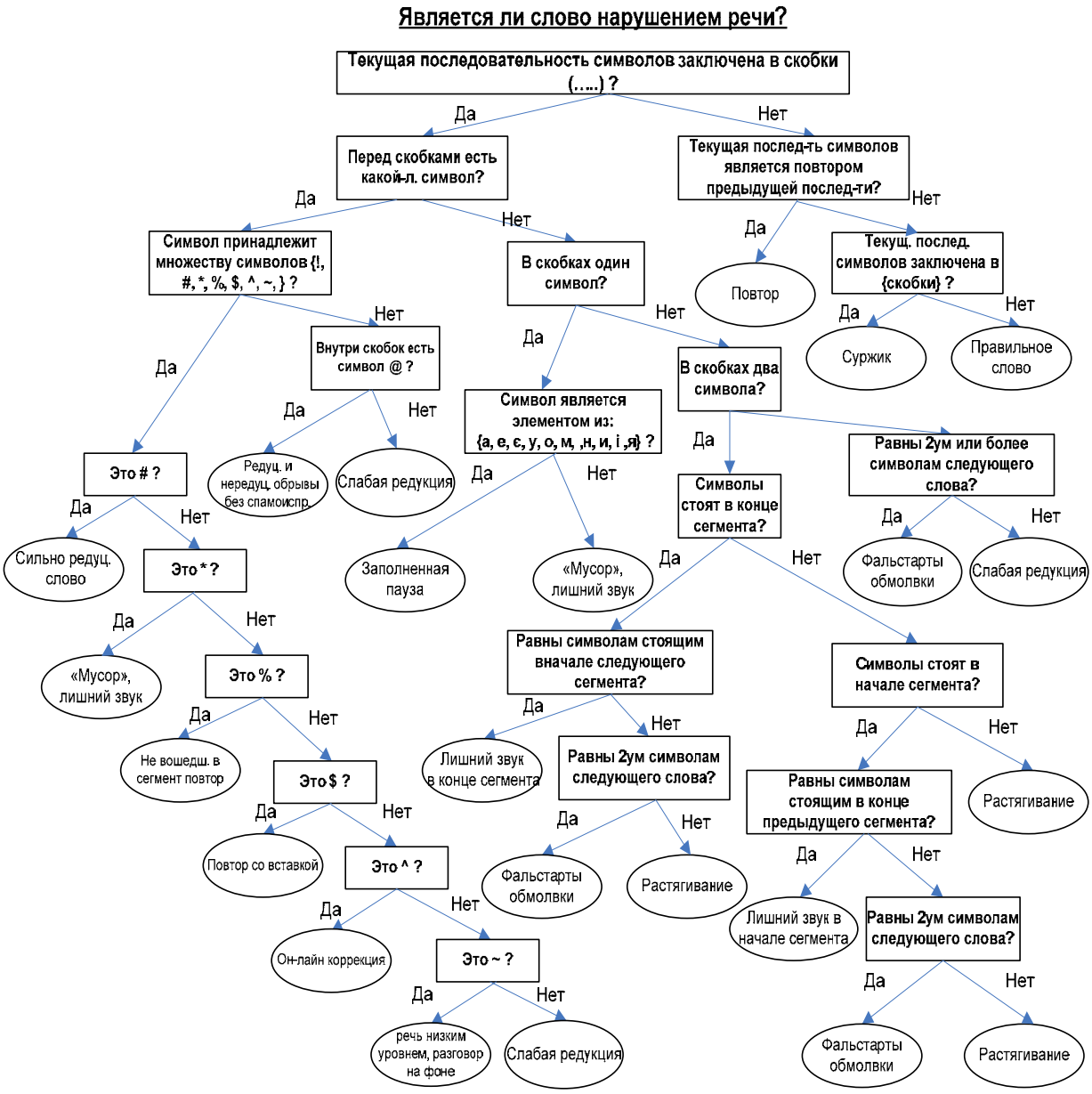

Рис. 2. Классификационное дерево нарушений речи

2.4 Правила автоматического выявления элементов спонтанной речи

Автоматическое обнаружение речевых сбоев начинается с выделения тех элементов речи, которые проще всего отделить от остальных элементов.

Выявление *заполненных* (*вокализованных*) *пауз* в нашем корпусе обеспечивается поиском ограниченного набора символов, состоящего из 10 элементов, по алгоритму:

If *word* ∈ [(а), (е), (є), (у), (о), (м), (н), (и), (і), (я)]
then
return "*вокализованная пауза*"
end if

Выявление *нефонологических* удлинений и слабо редуцированных слов осуществлялось по алгоритму:

If (*word = 2 tokens*) ∈ [(хе), (іе), (ее), (оа), (мм), …]
then
return "*растягивание*"
else
return "*редукция*"
end if

Обнаружение явлений, связанных со случайным попаданием в исследуемый сегмент лишних звуков (участков фонем) из соседних сегментов производилось по следующему алгоритму:

If ([word >=2 tokens] ∈ [end of string]) **and** ([word >=2 tokens] = [2 tokens of next string])
then
return "*неверная граница*"
else
If (word = 2 tokens) ∈ [(хе), (іе), (ее), (оа), (мм), …]
then
return "*растягивание*"
else
return "*редукция*"
end if
end if

Аналогичным образом производилось обнаружение лишних звуков в конце сегмента. Неинформативным элементам этого вида назначался вручную знак «*» – звездочка.

Правила определения *фальстартов и обмолвок* реализованы таким образом:

If getSimilarity(word >=2 token, nextWord>=2 token)
then
return "*фальстарт и обмолвка* "
end if

Повторы слов и словосочетаний выявлялись в текущем сегменте. Кроме того, вручную определялись повторы слов, вошедшие в разные сегменты. Для ускорения процесса поиска и исключения пропуска повторов производился дополнительный поиск всех слов, повторяющихся в текущем сегменте речи.

2.5. Эффективность расширенной системы разметки

Предварительная проверка эффективности расширенной системы разметки показала, что надежность АРСУР повышается в среднем на 1,25%, с наилучшим показателем 3,5% [10, 11]. При этом было замечено, что повышение надежности существенно зависит от дикторов.

Более тщательная разметка дополнительных материалов, представленных в работах [10, 11], позволила повысить надёжность АРСУР в среднем на 5,3…6,5%, с наилучшим показателем 7,62%.

Выводы

В работе произведено исследование звукового и текстового речевого материала заседаний Верховной Рады Украины, в котором общее количество нарушений спонтанной речи составило 4,9% от общего количества слов. Учитывая, что данный материал был предварительно подготовлен, - удалены шумы окружающей обстановки, вырезаны длительные паузы из речи говорящего, - можно утверждать, что число нарушений в спонтанной речи данного вида составляет не менее 5%.

Разработана расширенная система разметки явлений спонтанной украинской речи, необходимая для исследования влияния нарушений спонтанной речи на показатели надёжности системы АРСУР.

Экспериментально доказана правильность предположения о целесообразности разметки речевого материала в два этапа: на первом этапе стенограммы размечаются вручную, а на втором этапе производится автоматическая классификация нарушений спонтанной речи.

Предложены и детально описаны правила поиска нарушений спонтанной речи, необходимые для проведения автоматизированного анализа стенограмм и реализованные в виде специальной компьютерной программы.

Экспериментальные исследования свидетельствуют об эффективности расширенной системы разметки речи: надёжность АРСУР повышается в среднем на 5,3…6,5%.

Произведенный анализ нарушений спонтанной речи свидетельствует, что помимо учета и автоматического обнаружения явлений спон-

танной речи, еще одной важной задачей является разработка эффективных алгоритмов поиска границ сегментов спонтанной речи. Новые подходы к решению этой задачи должны облегчать деятельность эксперта и снижать количество ошибок сегментирования.

Полученные в данной работе результаты могут быть в дальнейшем использованы для формирования корпусов украинской речи.

Литература

1. *Zhao Y.,* Jurafsky D. A preliminary study of Mandarin filled pauses In Proc. of DiSS'05, Disfluency in Spontaneous Speech Workshop. 10–12 September 2005, Aix-en-Provence, France.
2. *Boula de Mareüil P.,* Habert B., Bénard F., Adda-Decker M., Barras C., Adda G., Paroubek P. A quantitative study of disfluencies in French broadcast interviews. In Proc. of DiSS'05, Disfluency in Spontaneous Speech Workshop. 10–12 September 2005, Aix-en-Provence, France.
3. *Moniz H ,* Trancoso I., Mata A. I. Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts, In Proc. Interspeech '2009, Brighton, U.K., 6-10 September 2009.
4. *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. – 2007. – №2. – С. 2–23;
5. *Богданова Н.В.* О корпусе текстов живой речи: новые поступления и первые результаты исследования // По материалам ежегодной международной конференции «Диалог» (2009) / Гл. ред. А.Е. Кибрик. М., 2009. С. 35–40.
6. *Лауринавичюте А.К.*, Федорова О.В. Влияние паузы хезитации на понимание синтаксической структуры предложения носителями русского языка // По материалам ежегодной международной конференции «Диалог» (2009) / Гл. ред. А.Е. Кибрик. М., 2009. С. 279–283.
7. *Леонтьева А.Б.*, Кипяткова И.С. Моделирование нефонемных речевых элементов и создание альтернативных транскрипций для распознавания спонтанной речи // Труды первого междисциплинарного семинара "Анализ разговорной русской речи" AR3 – 2007. – СПИИРАН. – Санкт-Петербург. – 2007. – С. 77–85;
8. *Лобанов Б.М.,* Цирульник Л.И. Моделирование внутрисловных и межсловных фонетико-акустических явлений полного и разговорного стилей в системе синтеза речи по тексту «Мультифон» // Труды первого междисциплинарного семинара "Анализ разговорной русской речи" AR3 – 2007. – СПИИРАН. – Санкт-Петербург. – 2007. – С. 57–71;
9. *Пилипенко В.В.,* Робейко В.В. Автоматизированный стенограф украинской речи. // Штучний інтелект. – Донецк – № 4. – 2008 р. – С. 768-775;
10. *Ладошко О.Н.,* Пилипенко В.В. Аннотация и учет речевых сбоев в задаче автоматического распознавания спонтанной украинской речи // Международная научно-техническая конференция «Искусственный интеллект. Интеллектуальные системы ИИ-2010», Тези доп., Том 1 – Донецк, 2010. – С. 223-227.
11. *Ладошко О.Н.,* Пилипенко В.В. Аннотация и учет речевых сбоев в задаче автоматического распознавания спонтанной украинской речи // Штучний інтелект. – Донецк – № 3. – 2010 р. – С. 238-248.

Национальный технический университет Украины «Киевский политехнический институт»