О распознавании фонем с помощью анализа речевого сигнала в частотной и временнòй областях.

Приложение к распознаванию синтаксически связных фраз

В.Ю. Шелепов,
доктор физико-математических наук

А.В. Ниценко, А.В. Жук, Д.С.Азаренко

Как известно, один из классических способов акустического описания фонем основан на использовании их формантной структуры — областей спектральных максимумов. В настоящей работе предлагается новый подход к этой проблеме, который сочетает в себе частотный анализ (обработку сигнала набором полосовых фильтров) с анализом профильтрованного сигнала во временнòй области.

Имея отрезок в 256 последовательных отсчётов записанного сигнала

$$x_1, x_2, \ldots, x_{256}$$

определим численный аналог полной вариации этого отрезка:

$$\sum_{i=1}^{256} \left| x_{i+1} - x_i \right| \qquad (1)$$

Если мы теперь выделим некоторый участок записанного речевого сигнала

$$x_1, x_2, \ldots, x_n, \ldots \qquad (2)$$

то его вариацией назовём среднее величин вида (1), вычисленных для последовательных окон в 256 отсчётов.

В работах [6], [9] подробно описан разработанный нами механизм автоматической сегментации речевого сигнала, т.е. разбиения его на участки, соответствующие отдельным фонемам с одновременным отнесением их к гласным, голосовым согласным, шипящим или паузообразным звукам. Основываясь на этом, мы выделяем участок записанного

речевого сигнала, соответствующий какой-либо одной фонеме. Затем обрабатываем сигнал полосовыми фильтрами с полосой пропускания от 0 до 200 Гц (получаем для него значения с плавающей точкой), умножаем профильтрованный сигнал на коэффициент 10 и вычисляем вариацию V_1 для выделенного участка профильтрованного сигнала. Затем обрабатываем сигнал фильтром с полосой пропускания от 200 до 400 Гц и вычисляем вариацию V_2 для выделенного участка 10-кратно увеличенного профильтрованного сигнала. Применяя таким же образом последовательно полосовые фильтры с полосой пропускания 200 Гц и заканчивая фильтром от 4800 Гц до 5000 Гц, получаем набор чисел

.

$V_1, V_2, ..., V_{25}$ Мы будем использовать также численный аналог полной вариации «с переменным верхним пределом» выделенного участка (2) записанного сигнала:

$$V(0) = 0, \qquad V(n) = \sum_{i=0}^{n-1} |x_{i+1} - x_i|$$

Определим также следующую величину. Пусть N_1 — максимальное число такое,

что $V(N_1) \le 255$. Далее полагаем

$$W(n) = V(n) \quad \text{при} \quad 0 \le n \le N_1,$$

$$W(N_1 + 1) = 0, \quad W(n) = \sum_{i=N_1+1}^{n-1} |x_{i+1} - x_i| \quad \text{при} \quad N_1 + 1 \le n \le N_2,$$

где N_2 — максимальное число такое, что и так далее. В результате возникает массив чисел.

Возьмём среднее этих чисел для выделенной части сигнала. Это среднее условимся называть «вариационной мерой» или просто «мерой» *M* выделенной части сигнала. Такая величина вычисляется для результатов фильтрации с упомянутыми выше полосами пропускания и получается набор чисел

$$M_1, M_2, ..., M_{25}$$

Наконец, вычисляются величины

$$Z_i = V_i / M_i, \quad i = 1,2,...,25. \qquad (3)$$

Мы исходим из представления о том, что фонема и слово — это акустически принципиально разные фонетические объекты. Фонема (и даже класс близких фонем) — объект спектрально сравнительно однородный, слово же, напротив, состоит из спектрально разнородных частей. Поэтому распознавая слова целиком, мы должны использовать тот или иной *вектор* признаков. Для распознавания же фонем (и их классов) более целесообразно использовать подходящий скалярный признак или набор независимых скалярных признаков, каждый из которых должен обеспечивать свой результат распознавания.

Речь — сложная система, которая в целом вполне детерминированна. В то же время общеизвестно, что любой мыслимый признак, который можно использовать при её распознавании, является случайной величиной. В этом нет противоречия,

подобные вещи происходят и в физике, когда при хаотическом движении отдельных молекул для их большого числа вырабатываются детерминированные макрохарактеристики, такие как температура, давление и т.д. Каждый признак следует рассматривать как случайную величину со своей функцией распределения, которая зависит от конкретного диктора, конкретного микрофона, конкретной звуковой карты. Последние два момента определяющие: какой смысл делать распознаватель инвариантным по отношению к диктору, если он зависит от микрофона? Мы считаем, что до тех пор, пока вопросы, связанные с независимостью от аппаратного обеспечения, не решены, целесообразнее разрабатывать быстро обучаемые системы распознавания речи с подстройкой под диктора. Аккуратное описание функции распределения — серьёзная задача, требующая большого статистического материала. В виду крайне ограниченной сферы применения (конкретный диктор, микрофон и т.д.) более или менее точное описание функции распределения становится нецелесообразным. В то же время, как правило, на основе нескольких примеров можно указать интервалы, куда чаще всего попадают значения признака для каждого члена рассматриваемой пары фонем. Значения за пределами этих интервалов разумно интерпретировать как отказ от распознавания.

Создавая обучаемую систему распознавания пары фонем, использующую один скалярный признак *X*, задаём два числа *a, b*. При

$$X < a \qquad (4)$$

считаем, что объект распознавания принадлежит первому классу, при

$$X > b \qquad (5)$$

— второму. При

$$a < X < b$$

не выполняется ни (4), ни (5), и мы имеем отказ от распознавания.

Вначале задаётся достаточно малое инициальное значение *a* и достаточно большое инициальное значение *b*. Если, пользуясь ими при распознавании, система не определит объект первого класса, то число *a* слишком мало. После того как пользователь укажет истинный результат, система должна заменить *a* вычисленным значением признака, увеличив последнее, скажем, на 0,1. Таким образом, в процессе обучения число *a* может только расти. Аналогично число *b* может только убывать. При этом обеспечивается все большая надёжность в случае принятия решения. Если, начиная с некоторого момента, окажется

$$a > b \qquad (6)$$

то при попадании *X* в *(b,a)* для распознаваемого объекта выполняются оба неравенства (4) и (5), т.е. он должен быть отнесён к обоим классам сразу, что невозможно, так как предполагается, что класс должен определяться однозначно. Таким образом, в случае (6) попадание *X* в *(b,a)* должно означать отказ от распознавания. Суммируя сказанное, получаем, что при

$$X < \min(a.b)$$

объект относится к первому классу, при

$$X > \max(a, b)$$

— ко второму классу, при

$$\min(a,b) < X < \max(a,b)$$

система отказывается от распознавания. Обучение состоит в модификации констант *a,b* и продолжается до тех пор, пока система не проработает без ошибок на протяжении, скажем, десяти циклов распознавания. Тогда распознаватель будет либо с высокой надёжностью принимать правильное решение, либо откажется от распознавания.

Теперь представим себе, что для полученной системы вероятность отказа от распознавания достаточно мала. Если мы для той же пары введём ещё несколько таких систем, использующих другие признаки, то, в соответствии со схемой независимых испытаний Бернулли, вероятность того, что все они одновременно будут отказываться от распознавания, станет существенно меньше. Все вместе построенные системы дадут желаемый распознаватель для рассматриваемой пары фонем, если в случае противоречия в результатах отдельных систем решение будет приниматься «по большинству голосов».

Запишем речевой сигнал, содержащий какую-либо пару гласных фонем, например, *А,И*, произнесённых как ударные (без редукции). Выделив участки сигнала, соответствующие этим фонемам, вычислим для них величины (3), которые обозначим через

$$Z_i(A) \quad \text{и} \quad Z_i(И), \qquad i = 1,2,...,25 \ . \qquad (7)$$

Для выделения компонент, лучше других различающих рассматриваемые фонемы, используем отношения величин (7). Пусть числа *k,l* таковы, что

$$Z_k(A)/Z_k(И) = \max_{1\le i\le 25}[Z_i(A)/Z_i(И)]$$

$$Z_l(A)/Z_l(И) = \min_{1\le i\le 25}[Z_i(A)/Z_i(И)] \ .$$

Тогда в качестве основного признака для различения *А, И* между собой возьмём величину

$$X(A,И) = Z_k / Z_l \ .$$

Настройка порогов для двухпорогового скалярного распознавателя описана выше. Для одного из авторов этой статьи и используемых им микрофона и звуковой карты при распознавании *А, И* между собой оказалось достаточно одного распознавателя описанного типа с параметрами *k=7,l=2, a=b=0,2*. Высокая надёжность распознавания в данном случае обеспечивается даже без заранее предусмотренного интервала отказа от распознавания.

Построим распознаватель такого типа для каждой пары следующего набора гласных:

А,Ё,И,О,У,Ы,W,Q (8)

Здесь символом *W* обозначено ударное *Е*, символом *Q* — ударное *Я*. Введение специальных обозначений для ударных *Е, Я* связано с тем, что только они имеют

достаточно определённое произношение. В безударном варианте они произносятся различными носителями языка по-разному. Для так называемой «младшей нормы» (более молодое поколение москвичей) они ближе к *И*, у сибиряков и в сценической речи — ближе к *Е*, *Я*. Для ряда пар не удаётся избежать случаев отказа от распознавания. В этом случае распознаватель выдаёт в качестве результата оба соответствующих символа. Если для какой-то пары вводятся дополнительные распознаватели, так, что общее число распознавателей для пары оказывается больше одного, то для них вычисляется совокупный результат «по большинству голосов». После всех попарных распознаваний результатом распознавания считается одна или несколько гласных, которые получились при распознаваниях пар максимальное число раз. Снабдив программу набором соответствующих флажков, мы получаем также возможность ограничивать распознавание лишь некоторыми гласными (8). Тогда остальные автоматически включаются в класс распознанных, поскольку на этом этапе для нас главное, чтобы слово «не потерялось», т.е. попало в формируемый программой список кандидатов на распознавание.

Таким же образом организуется распознавание в множестве голосовых согласных, множестве шипящих и аффрикат, множестве паузообразных звуков *К, П, Т, Ф, Х*. При распознавании в каждом из этих множеств в качестве результата в общем случае получается некоторый класс — часть этого множества фонем. Этот класс формируется автоматически. На *рисунке 1* представлен результат распознавания слова **«лиса».**

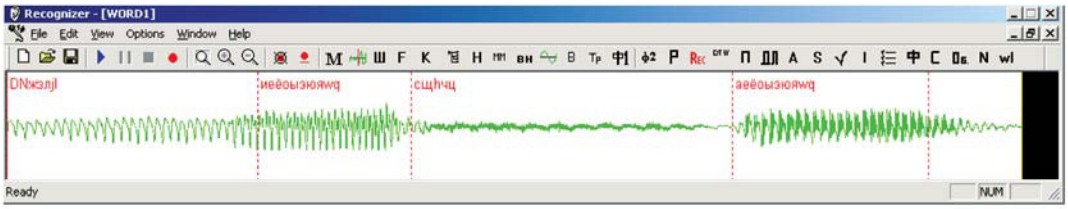

Рис. 1. Результат распознавания слова «лиса»

В работе [7] описан способ распознавания с предварительным заданием классов фонем, например, классов вида

аяqоёэеwуюыи W
аяqоёэеw Q
эewоёуюиыя Y
аяq A (9)
оёэеw R
эw E
оё O
.....

При этом распознаваемые цепочки из символов W,Q,Y,A,R… дают смешанную транскрипцию, состоящую из транскрипционных символов различного уровня детализации.

В идеале результатом распознавания фонем, образующих слово, служит его транскрипция, по которой слово в большинстве случаев однозначно восстанавливается. Однако любые признаки, используемые при распознавании речи, имеют характер случайных величин. Поэтому на любом этапе возможен отказ от распознавания и в результате вместо цепочки

транскрипционных знаков на выходе получается последовательность символов, обозначающих те или иные достаточно широкие классы фонем (смешение транскрипций разного уровня детализации). Возникает проблема, как по такому разнородному результату в большом словаре отыскать слова, которые ему удовлетворяют. В упомянутой работе [7] описан алгоритм, который основан на представлении транскрипций слов распознаваемого словаря в виде дерева и позволяет осуществить этот поиск очень быстро. В нашем случае распознанный класс фонем строится динамически. Введя для него виртуальное промежуточное обозначение типа (9), можно свести дело к алгоритму для заранее заданных классов (9). В результате получается способ поиска слов, соответствующих распознанной последовательности классов, причём длина поиска равна количеству распознанных классов, т.е. количеству фонем в сказанном слове. Это позволяет во времени, близком к реальному, находить списки слов-кандидатов на распознавание в случае словаря, содержащего миллионы словоформ.

Дальнейшее посвящено изложению концепции системы «речь — текст», позволяющей распознавать фразу как последовательность синтаксически связанных словоформ.

Известно, что в наиболее распространённом сейчас языке международного общения — английском — синтаксические связи между словами предложения осуществляются в основном с помощью предлогов. В результате за исключением некоторых моментов (типа «S» в конце слова в случае множественного числа), слово из текста всегда можно найти в орфографическом словаре. В отличие от английского русский язык относится к числу так называемых флективных языков. Большинство слов помимо начальной или словарной формы, называемой также «леммой», имеют достаточно развитую систему косвенных форм, образуемых с помощью флексий — частей, изменяемых при склонении, спряжении и т.д. Правильное использование этих форм — непременное условие синтаксически связной русской речи. Наличие многочисленных косвенных форм создаёт дополнительные трудности при компьютерном распознавании русской речи, ибо каждая из косвенных форм является для компьютера новым словом, в результате чего резко возрастает объём распознаваемого словаря. Различие же между отдельными формами часто сводится к безударным гласным в окончании, различать которые на сегодняшний день при обычной речи не представляется возможным. Последнее связано с редукцией упомянутых безударных гласных.

Имея дело с технической системой (хотя бы и относящейся к искусственному интеллекту), каковой является система компьютерного распознавания речи, мы вправе предложить пользователю соблюдать некоторые правила, относящиеся к самой речи. Например, можно на первых порах настаивать на подчёркнутой артикуляции, когда при слитном произнесении слова в нём слегка подчёркивается слоговая структура, так что каждая гласная становится как бы ударной.

Далее может быть предложена специальная архитектура системы, которая наряду с распознаванием речи использует элементы выбора из небольших словарей. Опишем одну из таких возможных систем. Распознаватель работает с первоначальным списком, содержащим все словоформы слов известного словаря Зализняка [10]. Получается достаточно полный словарь **всего русского язы-**

ка. Запись сказанного слова происходит при нажатии клавиши, соответствующей его первой букве, так что первая фонема при распознавании фактически заранее задаётся. В результате описанной выше процедуры пофонемного распознавания мы получаем список слов-кандидатов на распознавание — набор словоформ.

Далее используется наличие быстрого лемматизатора (система, восстанавливающая начальную форму слова по косвенной). Применяя к словам полученного списка лемматизатор, получим соответствующий набор начальных форм, который в несколько раз короче ранее полученного списка словоформ. По указанию пользователя (щелчок мыши на элементе списка) по исходному звуковому файлу строится эталон [5] и ему сопоставляется соответствующая лемма. Тогда, как показывает опыт, в подавляющем большинстве случаев при распознавании с использованием алгоритма DTW для любой словоформы этого слова построенный эталон окажется ближайшим и, следовательно, она будет отождествляться с указанной леммой. Исключения составляют ситуации типа «ИДТИ-ШЁЛ», «ЧЕЛОВЕК-ЛЮДИ». Далее компьютер отбирает из всех распознанных словоформ список словоформ, соответствующих распознанному слову и записанному сигналу. Это, собственно, и является результатом распознавания. В дальнейшем при произнесении других словоформ слова, для которого создан эталон, компьютер будет распознавать эти словоформы, используя эталон. Подчеркнём, что эталон строится по произвольной произнесённой словоформе, получает имя соответствующей леммы, и по нему распознаются (с точностью до леммы) все другие словоформы этого слова. Отметим также что, хотя дело заканчивается распознаванием по эталонам, применение пофонемного распознавания позволяет использовать эталоны в пределах списка распознанных словоформ, который на порядки меньше исходного полного словаря словоформ. На *рисунке 2* представлен результат распознавания слова **«солому»** по эталону, образованному по слову **«соломе»**.

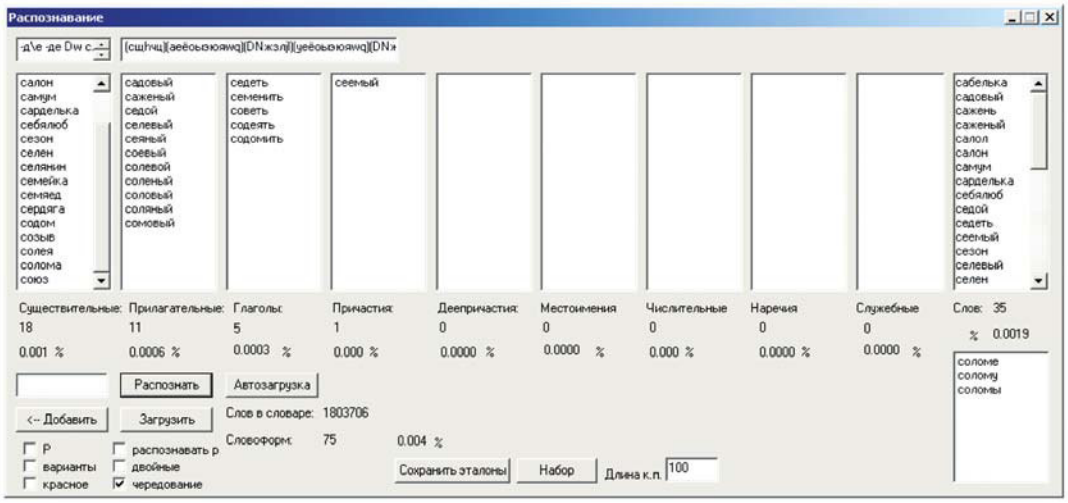

Рис. 2. Результат распознавания слова «солому» по эталону, образованному по слову «соломе»

В правом нижнем углу — список распознанных словоформ. При диктовке этот список передаётся в редактор, причём все его слова заключаются в скобки. Если список состоит только из одного слова, оно в скобки не заключается.

Пусть, например, диктуется фраза «Машина остановилась за углом». Тогда в окне редактора появятся следующие группы слов:

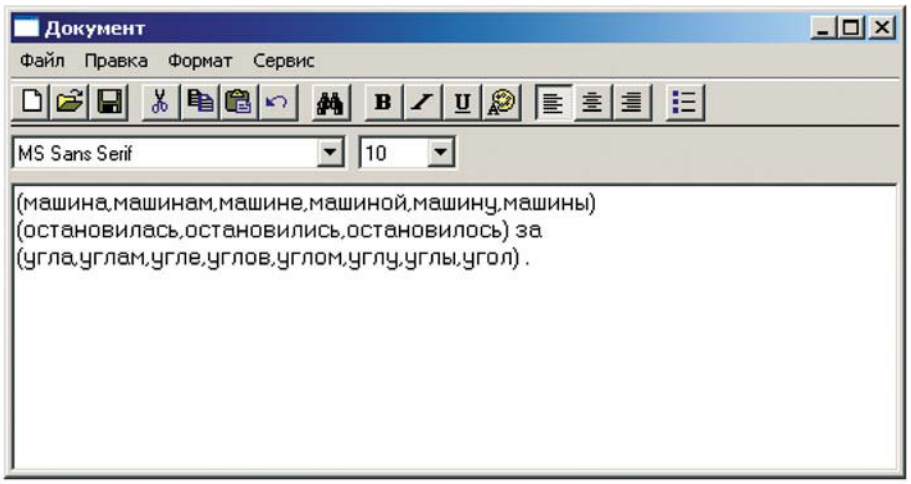

Рис. 3. Перечень групп при распознавании предложения

Далее, после того, как пользователь, закончив диктовать предложение, поставит точку, вступает в действие модуль синтаксической коррекции, работающий с использованием морфоанализатора, разработанного Г.В. Дорохиной [11, 12] и А.П. Павлюковой. Пользователь щелчком указывает подлежащее и выбирает для него нужную словоформу (в нашем примере — это слово «машина»). Компьютер убирает в отмеченной группе все лишние словоформы. Если при подлежащем есть прилагательное, в соответствующей группе автоматически оставляется только форма, согласующаяся с подлежащим. Аналогичным образом с подлежащим согласуется глагольное сказуемое. Среди форм существительного, следующего за предлогом, выбираются лишь те, которые этим предлогом допускаются. В нашем примере после первого шага получается следующее:

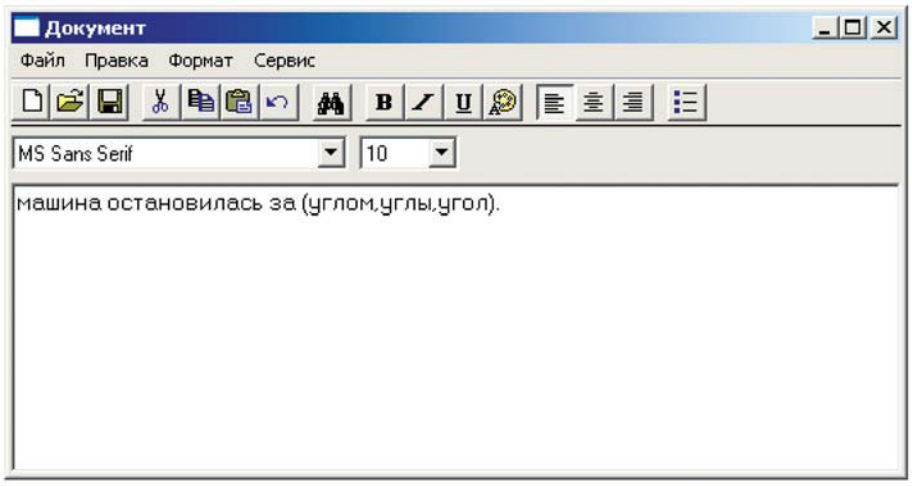

Рис. 4. Результат после выбора подлежащего и автоматического согласования с ним сказуемого

Конечный результат получается после выбора нужной словоформы в последней группе:

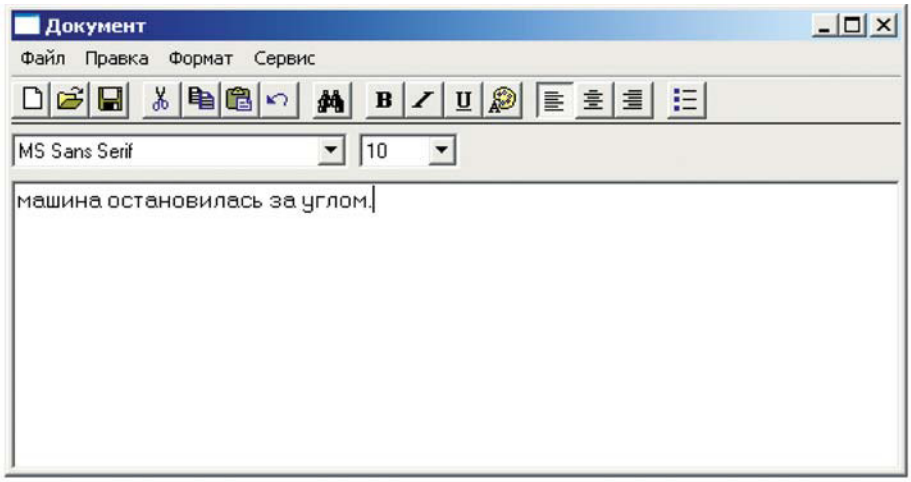

Рис 5. Конечный результат работы системы

Литература

1. *Дорохин О.А., Старушко Д.Г., Фёдоров Е.Е., Шелепов В.Ю.* Сегментация речевого сигнала // Искусственный интеллект. 2000. № 3. С. 450–458.
2. *Шелепов В.Ю., Ниценко А.В.* Амплитудная сегментация речевого сигнала, использующая фильтрацию и известный фонетический состав // Искусственный интеллект. 2003. № 3. С. 421–426.
3. *Ниценко А.В., Шелепов В.Ю.* Алгоритмы пофонемного распознавания слов наперёд заданного словаря // Искусственный интеллект. 2004. №. С. 633–639.
4. *Шелепов В.Ю., Ниценко А.В.* К проблеме пофонемного распознавания // Искусственный интеллект. 2005. № 4. С. 662–668.
5. *Засыпкин А.В., Мицевич А.Т., Овецкий М.В., Шелепов В.Ю.* О дикторонезависимой системе голосового телефонного номеронабирателя // Труды международной конференции «Знание-Диалог-Решение». Ялта. 1995. С. 427–430.
6. *Шелепов В.Ю., Ниценко А.В.* Структурная классификация слов русского языка. Новые алгоритмы сегментации речевого сигнала и распознавания некоторых классов фонем // Искусственный интеллект. 2007. № 1. С. 213–224.
7. *Шелепов В.Ю., Ниценко А.В., Жук А.В.* Новые алгоритмы распознавания фонем и их классов, поиск слова по его смешанной транскрипции при распознавании слов большого словаря // Искусственный интеллект. 2007. № 2. С. 139–147.
8. *Шелепов В.Ю., Ниценко А.В.* О распознавании фразы как последовательности синтаксически связанных словоформ // Искусственный интеллект. 2007. № 3. С. 344–346.
9. *Шелепов В.Ю.* Концепция пофонемного распознавания отдельно произносимых слов русской речи. Распознавание синтаксически связных фраз: Материалы международной научно-технической конференции Искусственный интеллект // Интеллектуальные системы (ИИ-2007). Донецк-Таганрог-Минск. С. 162–170.
10. *Зализняк А.А.* Грамматический словарь русского языка. М.: Русский язык. 1977.
11. Патент України № 78806 «Пристрій для збережёння і пошуку рядкових величин та спосіб збережёння і пошуку рядкових величин» Власник: Інститут проблем штучного інтелекту, винихідник Дорохіна Г.В. // Промислова власність. Бюл. № 5. 25.04.2007.
12. *Дорохина Г.В., Павлюкова А.П.* Модуль морфологического анализа слов русского языка // Искусственный интеллект. 2004. № 3. С. 636–642.

Шелепов Владислав Юрьевич —

*доктор физико-математических наук, профессор,
главный научный сотрудник Института искусственного интеллекта,
Начальник отдела распознавания речевых образов,
речью занимается с 1993 года.*

Ниценко Артём Владимирович —

младший научный сотрудник отдела распознавания речевых образов Института проблем искусственного интеллекта МОН и НАН Украины.

Жук Александр Викторович —

и.о. начальника отдела распознавания речевых образов Института проблем искусственного интеллекта МОН и НАН Украины.

Азаренко Дмитрий Сергеевич —

инженер отдела распознавания речевых образов Института проблем искусственного интеллекта МОН и НАН Украины.