7П

УДК 004.934

В.В. Пилипенко
Международный научно-учебный центр информационных технологий и систем,
г. Киев, Украина
valery_pylypenko@mail.ru

Распознавание дискретной и слитной речи из сверхбольших словарей на основе выборки информации из баз данных

Рассматривается двухпроходный алгоритм для распознавания дискретной и слитной речи из сверхбольших словарей (больше 1 млн) с использованием выборки информации из баз данных ELVIRCOS (Extra Large Vocabulary COntinuous Speech recognition based on the Information Retrieval). Процесс распознавания речи разделяется на два прохода, где первый проход используется для построения подмножества слов для второго прохода алгоритма. Представлен алгоритм формирования подсловаря на основе запросов к базе данных, а также процедура построения графа слов для слитной речи. Приведены результаты экспериментов по распознаванию дискретной и слитной речи с системой, содержащей практически все слова языка (около 2 млн слов).

Введение

Для распознавания речи из больших словарей для изолированно (дискретно) произносимых слов к началу 90-х годов было разработано несколько систем, которые показали достаточно высокие результаты [1]. Затем внимание исследователей переключилось на распознавание слитной речи, где статистические зависимости в порядке следования слов позволили предсказывать текущее распознаваемое слово на основе нескольких предыдущих слов. Это позволило разработать системы распознавания с суммарным объемом в десятки тысяч слов, хотя на каждом шаге распознавания рассматривалось несколько сотен альтернатив.

Тем не менее, существует необходимость построения систем распознавания речи со значительно большим количеством альтернатив при условии, что нет каких-либо ограничений на порядок распознаваемых слов.

Например, при управлении компьютера голосом невозможно предсказать следующее слово на основе нескольких предыдущих, поскольку это определяется логикой управления компьютером, а не свойствами текста. С другой стороны, существует необходимость значительного увеличения объема словаря для того, чтобы охватить все синонимы одной и той же команды, поскольку пользователю обычно трудно запомнить только один вариант названия команды.

Второй пример связан с диктовкой текстов. Использование подобных систем обычно ограничено такими текстами, которые аналогичны тем текстам, для которых накапливались статистики. К тому же дополнительное редактирование набранного текста требует наличия всех слов в активном словаре.

Таким образом, существуют приложения, где желательно иметь словарь максимально большого размера, чтобы в будущем охватить все слова для данного языка.

Дополнительная информация для ограничения числа альтернатив может быть получена из анализа речевого сигнала. Для этого предлагается выполнить пробное

распознавание при помощи так называемого фонетического стенографа. Полученная последовательность фонем формирует поток запросов к базе данных для выявления небольшого количества слов, которые могли бы входить в словарь распознавания.

Последующие разделы описывают новый двухпроходный алгоритм. Вначале представлена базовая система для сравнения с предложенной системой распознавания речи. Затем описаны два варианта алгоритма для изолированных слов и слитной речи и приведены результаты экспериментов.

1. Базовая система распознавания речи

Предложенный метод можно применить в любой системе распознавания речи, где представлены фонемы и можно сформировать процедуру фонетического стенографа. В данной работе как базовая система используется инструментарий HTK [2] на основе скрытых марковских моделей (HMM).

1.1. Предварительная обработка речевого сигнала

Речевой сигнал преобразуется в последовательность векторов признаков с интервалом анализа 25 мс и шагом анализа 10 мс. Вначале речевой сигнал фильтруется фильтром высоких частот с характеристикой $P(z)=1-0{,}97z^{-1}$ и применяется окно Хэмминга. Быстрое преобразование Фурье переводит временной сигнал в спектральный вид. Спектральные коэффициенты усредняются с использованием 26 треугольных окон, расположенных в мел-шкале. 12 кепстральных коэффициентов вычисляются при помощи обратного косинусного преобразования.

Логарифм энергии добавляется в качестве 13-го коэффициента. Эти 13 коэффициентов расширяются до 39-мерного вектора параметров путем дописывания первой и второй разностей от коэффициентов, соседних по времени. Для учета влияния канала применяется вычитание среднего кепстра.

1.2. Акустическая модель

Акустические модели отражают характеристики основных единиц распознавания. Для акустических моделей используются скрытые марковские модели с 64 смесями гауссовских функций плотности вероятности. 47 русских контекстно-независимых фонем моделируются тремя состояниями марковской цепи с пропусками.

Словарь транскрипций создается автоматически из орфографического словаря с использованием контекстно-зависимых правил.

1.3. Показатели базовой системы

Акустические модели обучались на выборке из 12 тыс. звуковых записей из словаря в 2037 слов и фраз, произнесенных одним диктором. Распознавание производилось на компьютере P-IV 2.4 ГГц.

Для проверки надежности распознавания речи было накоплено 1000 изолированных слов тем же диктором. Пословная надежность распознавания и среднее время распознавания одной секунды речи для различных размеров словаря приведены в табл. 1. Поскольку время распознавания линейно зависит от размера словаря, то для словаря в 1987 тыс. слов его можно оценить приблизительно в 2300 секунд.

Таблица 1 – Результаты распознавания изолированных (дискретных) слов базовой системой

Объем словаря, тыс.	1	15	95
Надежность, %	99,9	97,9	94,7
Время, с	1	16	115

Для проверки надежности распознавания слитной речи было дополнительно накоплено 1000 фраз с числами от 0 до 999. Пословная надежность распознавания и среднее время распознавания одной секунды речи для различных размеров словаря приведены в табл. 2.

Таблица 2 – Результаты распознавания слитной речи базовой системой

Объем словаря, тыс.	1	15	95
Надежность, %	98,0	96,5	92,6
Время, с	2,1	36	205

2. Алгоритм ELVIRS для изолированно (дискретно) произносимых слов

2.1. Архитектура

Архитектура системы распознавания ELVIRS (Extra Large Vocabulary Speech recognition based on the Information Retrieval) показана на рис. 1. Такие блоки из базовой системы как *вычисление признаков* и *акустических моделей* используются перед первым проходом алгоритма. Также на втором проходе используется обычное *сравнение образов* в условиях ограниченного словаря.

Изменения касаются дополнительного первого прохода алгоритма, где *фонетический стенограф* используется для получения последовательности фонем. Затем процедура *выборки информации* создает ограниченный словарь для второго прохода алгоритма.

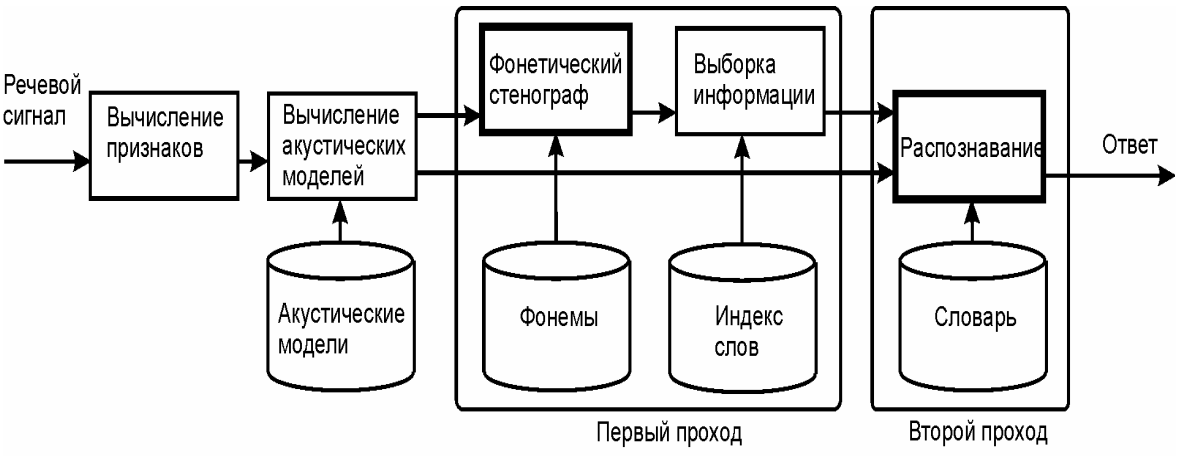

Рисунок 1 – Архитектура двухпроходной системы распознавания речи из сверхбольших словарей

2.2. Фонетический стенограф

Алгоритм фонетического стенографа [3], [4] позволяет строить фонетическую последовательность для речевого сигнала без использования какого-либо словаря. Для этой цели строится некоторая генеративная грамматика, которая может синтезировать все возможные модельные сигналы непрерывной речи для любой последовательности фонем. В рамках построенной модели строится алгоритм пофонемного распознавания для неизвестного сигнала.

Используются те же контекстно-независимые модели фонем, что и в базовой системе распознавания.

Граф процесса распознавания фонем показан на рис. 2. Следует заметить, что на процесс распознавания накладываются гораздо менее сильные ограничения, чем при распознавании слов, задаваемыми транскрипциями. Это позволяет распознавать произнесенный сигнал независимо от словаря.

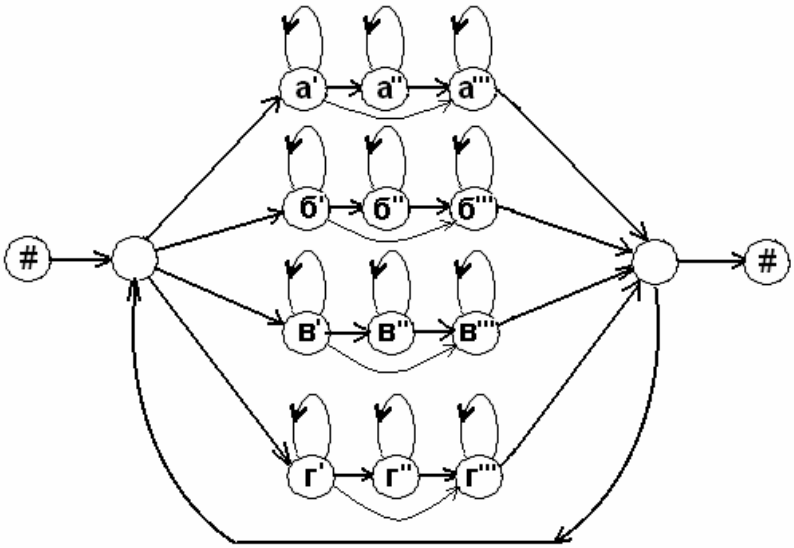

Рисунок 2 – Граф процесса распознавания произвольной последовательности фонем

В табл. 3 приведены показатели надежности найти фонему на правильном месте для известной реализации.

Таблица 3 – Надежность распознавания последовательности фонем

	Выборка	Надежность, %
1	Va0	85,2
2	Va1	85,0
3	Va4	85,8
4	Vk0	84,1
5	Vk1	83,2
6	Vz0	83,4

Каждая выборка содержала по 1000 слов. Выборки Va0 и Va1 входили в обучающую выборку. Выборка Va4 состоит из слов, которые входили в ОВ, но были произнесены дополнительно. Выборки Vk0, Vk1, Vz0 состоят из слов, которые не входили в ОВ.

Результаты показывают небольшую надежность распознавания фонем, поскольку для распознавания слова или фразы необходимо, чтобы все фонемы в последовательности были распознаны правильно.

2.3. Процедура получения подсловаря из базы данных

Заранее в процессе обучения из словаря транскрипций создается индекс от троек фонем к транскрипциям. Ключом индекса является тройка фонем. Таким образом, таблица индекса состоит из M^3 вхождений, где M есть число фонем в системе. Каждое вхождение в таблицу содержит список транскрипций, в которые входит тройка фонем ключа вхождения.

Нетрудно посчитать дополнительные затраты памяти по сравнению с базовым алгоритмом распознавания речи. Каждая тройка фонем в транскрипции порождает одну ссылку, занимающую одну ячейку памяти, на номер транскрипции, из которой взята данная тройка фонем. Суммарное число данных в индексе равно числу всех троек фонем во всем словаре транскрипций. Число троек в транскрипции i-го слова равно длине транскрипции L_i – 2. Таким образом, дополнительные затраты памяти по сравнению с базовым алгоритмом распознавания речи равны $M^3 + \sum_{i=1}^{K}(L_i - 2)$ ячеек памяти, где K – число транскрипций в словаре. Например, для словаря в $2M$ слов это составляет около 50 Мб, что вполне доступно для современных компьютеров.

Процесс получения подсловаря иллюстрируется на рис. 3. Выход фонетического стенографа делится на пересекающиеся тройки фонем со сдвигом в одну фонему. Получающаяся тройка фонем становится запросом к базе данных. Сейчас в настоящей версии используется простой запрос, когда он в точности совпадает с тройкой фонем. В будущем предлагается использовать расстояние *Levensteine* для учета вставок, удалений и замен в последовательности фонем. Таким образом, последовательность фонем продуцирует поток запросов к базе данных.

Ответ на один запрос состоит из списка транскрипций, в которые входит данная тройка фонем. Этот список копируется в подсловарь для второго прохода алгоритма. Следующий запрос из потока добавляет новую порцию транскрипций, при этом подсчитывается количество повторений для того, чтобы можно было вычислить ранг слова в подсловаре.

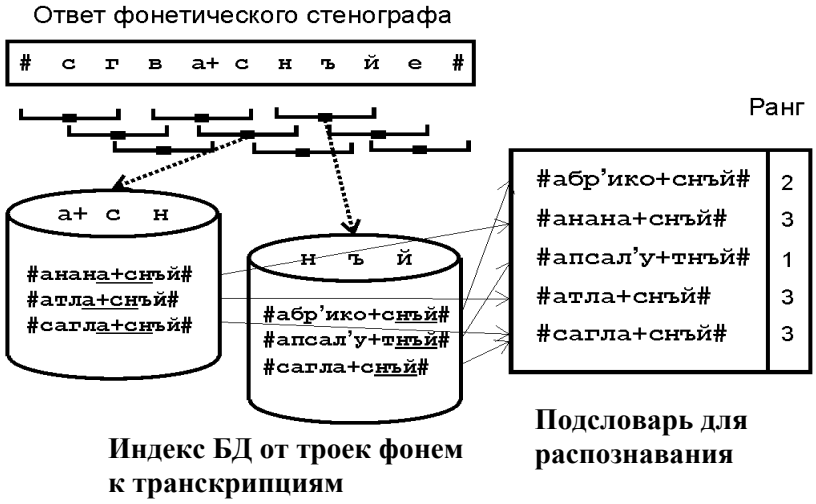

Рисунок 3 – Процесс получения подсловаря

Все транскрипции в полученном подсловаре упорядочиваются согласно рангу слова (счетчику повторений). Первые N транскрипций заносятся в окончательный подсловарь для второго прохода алгоритма. Таким образом, подсловарь для распознавания содержит транскрипции с наивысшими рангами и число транскрипций не превышает фиксированного числа N.

2.4. Алгоритм ELVIRS распознавания изолированных слов

Алгоритм ELVIRS [5] состоит из двух частей.

Подготовительный этап:

1. Подготовить словарь для распознавания.
2. Выбрать множество фонем и создать транскрипции слов из словаря при помощи правил.
3. Создать индекс базы данных от троек фонем к транскрипциям.
4. Обучить акустические модели по накопленным речевым сигналам.

Этап распознавания:

1. Применить фонетический стенограф к входному сигналу для получения последовательности фонем.
2. Поделить последовательность фонем на пересекающиеся тройки фонем со сдвигом в одну фонему.
3. Создать запросы к БД из троек фонем.
4. Получить списки транскрипций при помощи запросов к индексу базы данных.
5. Ранжировать транскрипции по их рангу.
6. Выбрать первые N транскрипций с наивысшими рангами в качестве подсловаря для распознавания.
7. Распознать входной речевой сигнал в условиях ограниченного подсловаря.

3. Информационная оценка вероятности правильного формирования подсловаря

Ответ распознавания фонетического стенографа может рассматриваться как правильная последовательность фонем, пропущенная через канал с шумом. Обозначим в ответе фонетического стенографа правильную фонему как 1, а испорченную шумом как 0. В табл. 4 первая строка показывает правильную транскрипцию слова, во второй строке приведен ответ фонетического стенографа.

Таблица 4 – Пример кодирования ответа фонетического стенографа

#	с	а	г	л	а+	с	н	ъ	й	#
#	с	т	г	в	а	с	н	ъ	й	#
1	1	0	1	0	0	1	1	1	1	1

Пусть вероятность появления 1 в двоичном наборе равна u. Вероятность P найти в двоичном наборе длины n подряд k единиц и больше можно вычислить при помощи следующего рекуррентного выражения:

$$P_n = \begin{cases} 0, & n < k \\ u^k, & n = k \\ P_{n-1} + u^k (1-u)(1-P_{n-k-1}), & n > k \end{cases}.$$

В табл. 5 показаны вероятности P найти в двоичных наборах подряд три и больше 1 при некоторых длинах n и вероятностях u. Средняя длина транскрипций равна приблизительно 8 и вероятность правильного нахождения фонемы в известных реализаций приблизительно равна 85 %. Для таких значений вероятность найти правильное слово в подсловаре равна 0,953.

Таблица 5 – Вероятность найти подряд три и больше 1 в двоичном наборе длины n

u / n	0,75	0,8	0,85	0,9
3	0,422	0,512	0,614	0,729
4	0,527	0,614	0,706	0,802
5	0,632	0,716	0,798	0,875
6	0,738	0,819	0,890	0,948
7	0,799	0,869	0,926	0,967
8	0,849	0,908	**0,953**	0,982
9	0,887	0,937	0,971	0,991
10	0,915	0,956	0,981	0,995

4. Алгоритм Elvircos для распознавания слитной речи

4.1. Архитектура

После получения списков транскрипций используется дополнительная процедура *формирование графа слов* для слитной речи, которая создает сеть слов для второго прохода алгоритма.

4.2. Формирование графа слов

Процесс создания графа слов показан на рис. 4. Сеть слов начинается из вершины S и заканчивается в вершине F. Каждая тройка фонем из ответа фонетического стенографа порождает промежуточную вершину с номером, синхронным времени появления тройки фонем. С другой стороны, каждая тройка фонем становится запросом к индексу базы данных, который возвращает список транскрипций. Транскрипции вставляются между промежуточными вершинами так, чтобы порождающие тройки фонем оказались в одной колонке по вертикали.

В случае, когда происходит пересечение транскрипций одного слова, порожденных разными тройками фонем, тогда ранги этих транскрипций увеличиваются на единицу. Для каждого момента времени можно подсчитать число транскрипций, входящих в этот момент времени.

Для уменьшения сложности графа слов используется ограничение N для количества слов в каждый момент времени. При этом удаляются слова с малыми рангами.

Поскольку граф слов формируется слева направо, можно производить его формирование в реальном времени с задержкой, равной максимальной длине транскрипции.

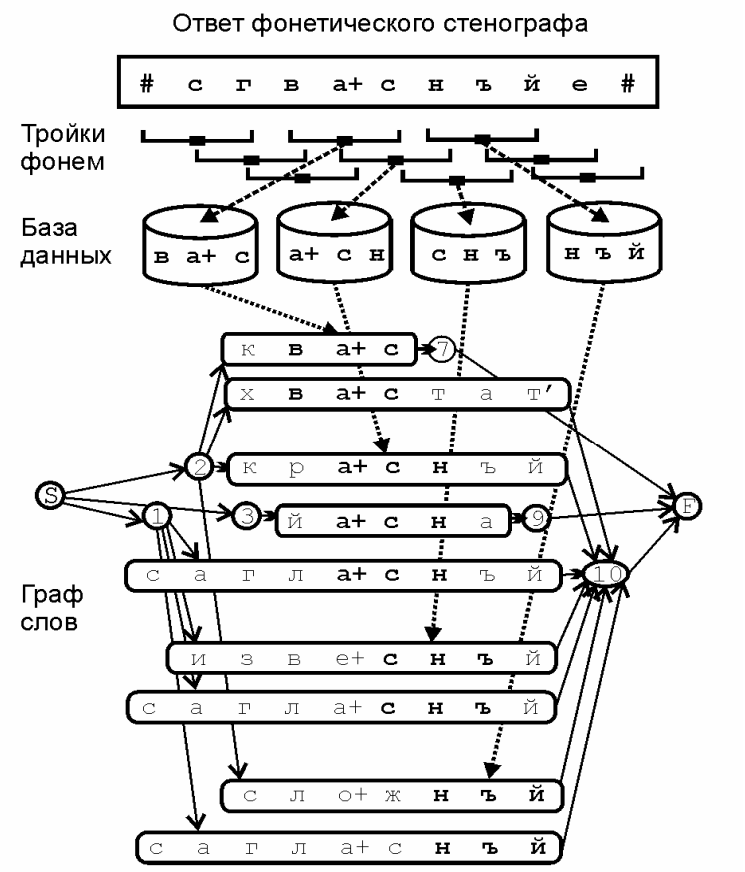

Рисунок 4 – Формирование графа слов для слитной речи

4.3. Алгоритм ELVIRCOS распознавания слитной речи

Алгоритм ELVIRCOS состоит из двух частей.

Подготовительный этап:

1. Подготовить словарь для распознавания.
2. Выбрать множество фонем и создать транскрипции слов из словаря при помощи правил.
3. Создать индекс базы данных от троек фонем к транскрипциям.
4. Обучить акустические модели по накопленным речевым сигналам.

Этап распознавания:

1. Применить фонетический стенограф к входному сигналу для получения последовательности фонем.
2. Поделить последовательность фонем на пересекающиеся тройки фонем со сдвигом в одну фонему.
3. Создать запросы к БД из троек фонем.
4. Получить списки транскрипций при помощи запросов к индексу базы данных.
5. Создать граф слов для слитной речи.
6. Ранжировать транскрипции по их рангу.
7. Выбрать первые N транскрипций с наивысшими рангами в качестве подсловаря для распознавания.
8. Распознать входной речевой сигнал для графа слитной речи в условиях ограниченного подсловаря.

5. Экспериментальные результаты

Для того чтобы ввести первый проход алгоритма ELVIRCOS в базовую систему распознавания речи, были сделаны необходимые изменения в инструментарии HTK и проведены несколько экспериментов.

Для изолированно произносимых слов исследовалось влияние ограничения N на среднее время и надежность распознавания для различных объемов словарей, что показано в табл. 6. Эксперименты проводились на тех же самых контрольных выборках, что и базовой системе. Результаты показывают полезность ограничения N для больших объемов словарей, что позволяет дополнительно сократить время распознавания при незначительном ухудшении надежности.

В целом получено значительное сокращение времени распознавания в сотни раз при относительно небольшом (около 5 %) ухудшении надежности по сравнению с базовой системой распознавания. Ухудшение надежности имеет хорошее совпадение с информационной оценкой вероятности правильного формирования подсловаря.

Таблица 6 – Результаты распознавания для алгоритма ELVIRS

Объем словаря, тыс. / Ограничение на размер подсловаря N	15		95		1987	
	Надежность, %	Время, с	Надежность, %	Время, с	Надежность, %	Время, с
50	92,2	1,4	81,0	1,4	69,2	1,6
200	94,6	1,6	87,6	2,1	76,0	1,9
500	**95,5**	**1,9**	**90,1**	**2,5**	80,0	3,3
1000	96,0	2,1	90,7	3,1	82,7	4,4
2000	96,0	4,4	92,0	4,5	**84,8**	**6,8**
5000	96,0	4,6	92,9	8,3	86,4	12,0

Для слитной речи были проведены предварительные эксперименты, в которых рассматривался случай, когда ограничение N было равно размеру словаря. В табл. 7 представлены показатели надежности для различных размеров словаря.

Таблица 7 – Результаты распознавания для алгоритма ELVIRCOS

Объем словаря, тыс.	15	95	1987
Надежность, %	85,3	84,9	83
Время, с	2,3	9,4	160

Уменьшение среднего времени распознавания не настолько значительное, как в случае изолированно произносимых слов, поскольку некоторые последовательности фонем от фонетического стенографа порождают графы слов очень большой размерности. Скорее всего, введение ограничения N существенно уменьшит время распознавания.

Выводы

Статья описывает новый подход к распознаванию речи из больших словарей и представляет экспериментальную проверку предложенных подходов. Следует подчеркнуть важность подходов выборки информации из баз данных для процесса распознавания речи. Показано, что дополнительный источник информации, полученный из анализа последовательности фонем от фонетического стенографа, позволил значительно сократить пространство поиска в алгоритме распознавания речи. Это позволило создать системы распознавания речи, охватывающие практически все слова языка.

Литература

1. Bahl L., Brown P., De Souza P., Mercer R., and Picheny M. Acoustic Markov models used in the tangora speech recognition system // Proc. ICASSP'88. – NewYork, NY. – 1988.
2. The HTK Book / S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland. – Cambridge University Engineering Department, 2002.
3. Taras K. Vintsiuk. Generative Phoneme-Threephone Model for ASR // Proc. of the 4th Workshop on Text, Speech, Dialog – TSD'2001. – Zelezna Ruda (Czech Republic). – 2001. – P. 201.
4. Пилипенко В.В. Використання фонетичного стенографа при розпізнаванні мовлення з великих словників // Тезисы 12-й Междунар. конф. «Автоматика – 2005». – Харьков. – 2005. – С. 73.
5. Пилипенко В.В. Технология распознавания большого количества образов на примере распознавания речи из сверхбольших словарей // Искусственный интеллект. – 2006. – № 2. – С. 332-335.

В.В. Пилипенко

Розпізнавання дискретного та злитого мовлення з надвеликих словників із застосуванням вибірки інформації з баз даних

Розглядається двопрохідний алгоритм розпізнавання дискретного та злитого мовлення з надвеликих словників (більш ніж 2 млн) із застосуванням вибірки інформації з баз даних ELVIRCOS (Extra Large Vocabulary COntinuous Speech recognition based on the Information Retrieval). Процес розпізнавання мовлення розподіляється на два проходи, де перший прохід формує підмножину слів для другого проходу алгоритму. Представлено алгоритм формування підсловника за допомогою запитів до бази даних, а також процедуру побудови графа слів для злитого мовлення. Наведені результати експериментів по розпізнаванню дискретного та злитого мовлення із системою, що включає практично всі слова мови (приблизно 2 млн слів).

V.V. Pylypenko

Discrete and Continuous Speech Recognition for Extra Large Vocabulary Based on Information Retrieval from Data Base

In this paper a new two-pass algorithm for Extra Large (more than 1M words) Vocabulary COntinuous Speech recognition based on the Information Retrieval (ELVIRCOS) is presented. The principle of this approach is to decompose a recognition process into two passes where the first pass builds the word subset for the second pass recognition. Sub-vocabulary retrieval procedure and word graph composition for continuous speech is presented. Experimental results for speech recognition system with vocabulary of about all words (approximately 2 M) are presented.

Статья поступила в редакцию 29.06.2006.