УДК 004.522

Д.А. Важенина, И.С. Кипяткова, К. Марков, А.А. Карпов

МЕТОДИКА ВЫБОРА ФОНЕМНОГО НАБОРА ДЛЯ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ

Важенина Д.А., Кипяткова И.С., Марков К., Карпов А.А. **Методика выбора фонемного набора для автоматического распознавания русской речи.**

Аннотация. В статье описывается выбор оптимального фонемного набора для системы автоматического распознавания русской речи. При создании акустических моделей был предложен комбинированный метод для выбора наилучшего фонемного набора, объединяющий статистическую информацию и фонетические знания. В результате применения данного метода к русскому фонетическому набору алфавита IPA (International Phonetic Alphabet) был получен набор из 47 фонологических единиц, который был преобразован в несколько фонемных наборов с разным размером от 27 до 47 единиц. Эксперименты по распознаванию речи показали, что использование сокращенных фонемных наборов позволяет увеличить точность распознавания фонем. В ходе экспериментов с применением расширенной языковой модели и сверхбольшим словарем точность распознавания слов составила 73,1%. Полученные результаты соответствуют качеству распознавания слитной русской речи, полученному на настоящий момент другими организациями.

Ключевые слова: автоматическое распознавание русской речи, акустическое моделирование, выбор фонемного набора.

Vazhenina D.A., Kipyatkova I.S., Markov K., Karpov A.A. **Technique for Phoneme Set Selection for Automatic Russian Speech Recognition.**

Abstract. In the paper, selection of best phoneme set for Russian automatic speech recognition is described. For the acoustic modeling, we describe a method based on combination of knowledge-based and statistical approaches to create several different phoneme sets. Applying this method to the Russian phonetic set of the IPA (International Phonetic Alphabet) alphabet, we first reduced it to 47 phonological units and derived several other phoneme sets with different number of phonological units from 27 till 47. Speech recognition experiments using these sets showed that reduced phoneme sets are better for phoneme recognition task and as good for word level speech recognition. For experiment with extra-large vocabulary, we used syntactico-statistical language model, which allowed us to achieve the word recognition accuracy of 73.1%. The results correspond to continuous Russian speech recognition quality obtained by other organizations up to date.

Keywords: automatic Russian speech recognition, acoustic modeling, phoneme set selection.

1. Введение. Автоматическое распознавание слитной русской речи представляет собой очень сложную задачу из-за ряда особенностей языка. Русский язык является флективным языком со сложной системой словообразования, что приводит к существенному увеличению размера словаря системы распознавания, а также к увеличению коэффициента неопределенности (perplexity) статистических *n*-граммных моделей языка [1]. Для того чтобы создать словарь большого объема, обычно используются автоматические транскрипторы, которые преобразуют слова в фонематическую транскрипцию, используя правила транскрибирования. Для русского языка правила транскриби-

рования не являются сложными, основная проблема состоит в том, чтобы определить положение ударения в слове, кроме того, в сложных словах может быть несколько ударных гласных [2].

При создании фонематических транскрипций обычно используется Международный фонетический алфавит (International Phonetic Alphabet) IPA/МФА, русский вариант которого состоит из 55 фонологических единиц: 38 согласных и 17 гласных. Большое количество согласных, которые более сложны для распознавания за счет меньшей длительности звучания, обусловлено специфичной для русского языка палатализацией большинства твердых согласных. Такие пары имеют небольшое различие в дополнительной артикуляции: подъем средней части языка к небу во время основной артикуляции согласного. Наличие таких пар согласных усложняет распознавание, поскольку увеличивает схожесть согласных между собой. Для сравнения: в американском варианте английского алфавита IPA содержится 49 фонологических единиц: 24 согласных и 25 гласных и дифтонгов. Таким образом, при меньшем общем числе фонем количество согласных значительно меньше, чем в IPA для русского языка, и их соотношение к гласным и дифтонгам, которые менее сложны для распознавания, примерно равное.

Набор гласных фонем в алфавите IPA включает также их редуцированные варианты. Все безударные гласные в русском языке подвержены редукции в длительности и все, кроме [u], подвержены артикуляторной редукции. Так безударный гласный звук [e] становится близким по произношению к [i], а безударный гласный [o] в большинстве случаев произносится как [a], исключая случаи с иностранными словами, такими как “радио”.

Выбор фонемного набора является одним из первых шагов при разработке системы распознавания речи. Под фонемным набором подразумевается совокупность фонологических единиц, которые используются как базовые акустические единицы при создании акустических моделей в системе распознавания речи. От выбора этого набора может зависеть точность и быстродействие всей системы.

В данной статье представлено описание методики выбора наилучшего набора фонем для систем распознавания русской речи. Предложенная методика основана на использовании как фонетических знаний, так и статистической информации. Для оценки эффективности методики были проведены эксперименты с использованием контекстно-независимых моделей для распознавания фонем и контекстно-зависимых моделей для распознавания слов.

2. Обзор систем распознавания русской речи. Обзор систем автоматического распознавания русской речи приведен в [3-5]. Ниже рассмотрены некоторые новые работы в этой области.

В [6] описывается система распознавания русской речи, разработанная в ходе проекта Quaero (http://www.quaero.com/). Система использует два различных подхода для обучения акустических моделей. 4-граммная модель языка со словарем 500 тыс. слов была обучена на материалах радионовостей, интернет-данных, книгах, транскрипциях аудиоданных. Точность распознавания слов, полученная на корпусе Quaero 2010, составила 80 %.

В работе [7] представлена модель русского языка на основе максимума энтропии, использующая признаки, которые позволяют учесть такие особенности русского языка как флективность и свободный порядок слов. Эта модель, объединенная с моделью языка, основанной на частях слов, использовалась для переоценки списка лучших N гипотез, что позволило увеличить точность распознавания слов на 1,2 %.

Система распознавания слитной русской речи с большим словарем, использующая слоговую модель языка, представлена в [8]. Словарь системы содержит 12 тыс. слогов. В работе предложен метод объединения слов и коррекции ошибок. Гипотеза произнесенной фразы создается из распознанных символов путем применения коэволюционного асимптотического вероятностного генетического алгоритма CAPGA (co-evolutionary asymptotic probabilistic genetic algorithm).

В [9] описывается метод учета синтаксических связей в модели языка, при котором используются следующие стадии обработки обучающего текстового корпуса: обозначение части речи, анализ синтаксических зависимостей и создание факторной модели языка для переоценки гипотез распознавания. Наилучшая точность распознавания, полученная при проведении экспериментов на части Национального корпуса русского языка, составила 91,77 %, что на 1,26 % лучше, чем точность, полученная при применении базовой модели.

Распознавание слитной русской речи с применение глубоких нейронных сетей (deep neural networks) совместно со скрытыми марковскими моделями представлено в [10]. На первой стадии распознавания для вычисления векторов признаков использовались глубокие нейронные сети. На второй стадии декодер Витерби осуществлял генерацию распознанной последовательности слов, используя вероятности, полученные на первой стадии. Эксперименты были выполнены на корпусе телефонной речи, содержащем 25 часов обучающих данных, 1 час данных для настройки системы и 1 час тестовых данных. Дополнительно 17 часов неразмеченных речевых данных использовались для

предварительного обучения глубоких нейронных сетей. Наивысшая точность распознавания составила 45 %, при этом использовалась следующая конфигурация нейронных сетей: 5 слоев с 1000 элементами на один слой.

Для голосового поиска в сети Интернет компания Google разработала функцию Voice Search [11], которая позволяет пользователям находить нужную им информацию, произнося слова или фразы вслух. Для создания модели языка использовались запросы пользователей к поисковой системе Google. Также эта технология применяется для других сервисов Google, например, она позволяет использовать речевые запросы для поиска места на карте.

Недавно компания Яндекс представила свою систему распознавания речи SpeechKit [12]. В настоящее время данное приложение позволяет искать общую информацию (то, что люди обычно ищут в Интернете) и геоинформацию (адреса, названия организаций). Точность распознавания общей информации составляет 84%, геоинформации – 94%.

В работе [13] представлена система автоматического создания в режиме реального времени субтитров для телевизионных новостей. Система работает следующим образом: специально обученный диктор (респикер) пересказывает речь диктора новостей, преобразуя ее в форму, более соответствующую письменной речи. Система распознавания речи обучена на голос респикера, что позволяет увеличить точность распознавания (точность распознавания слов составила 94,5 %). Затем редактор субтитров исправляет ошибки, допущенные системой распознавания. Таким образом, количество точность распознавания слов в субтитрах после их проверки редактором составляет 99,98%. Данная система использовалась в ходе паралимпийских игр в Сочи.

3. Акустическое моделирование и выбор фонемного алфавита. Создание моделей акустических единиц речи является первым этапом обучения системы автоматического распознавания речи. В зависимости от задачи и объема обучающих данных необходимо определить список моделируемых фонетических единиц, в качестве которых могут использоваться части слова, слоги, контекстно-независимые фонемы или контекстно-зависимые фонемные реализации. Преимуществом использования контекстно-зависимых единиц является их способность моделировать эффекты коартикуляции между соседними звуками, поэтому в современных системах распознавания речи контекстно-независимые модели (монофоны), которые соответствуют фонологическим единицам фонемного набора, часто заменяются контекстно-зависимыми моделями (трифонами). Как видно из примера, приведен-

ного в таблице 1, монофон /a/ встречается в модели дважды, но, с учетом контекста левой и правой фонем, они образуют разные трифоны.

Таблица 1. Фонетическое моделирование слова "пара" с использованием контекстно-независимых и контекстно-зависимых моделей

	Фонетические единицы
Монофоны	/p/ /a/ /r/ /a/
Трифоны	/sil-p+a/ /p-a+r/ /a-r+a/ /r-a+sil/

3.1. Разработка дерева решений для русского языка. При использовании трифонов может возникнуть проблема дефицита обучающих данных. Для обучения каждой модели трифона необходимо наличие большого числа наблюдений соответствующего трифона, и зачастую не все возможные трифоны присутствуют в обучающих данных. Наиболее распространенным решением данной проблемы является кластеризация и связывание состояний моделей трифонов, чьи контексты относятся к одному кластеру. Обычно кластеризация контекстов осуществляется с помощью деревьев решений [14,15], при этом используются вопросы о том, имеет ли левая или правая фонема определенные фонетические признаки (например, является ли левая/правая фонема звонкой).

В данном исследовании в качестве основы для разработки фонетического дерева для русского языка было взято фонетическое дерево для английского языка [16], которое было дополнено вопросами, относящимися к специфике фонологии русского языка, например, является ли левая/правая контекстная фонема мягкой. Некоторые вопросы, которые относятся к особенностям английского языка, были удалены. В таблице 2 приведены основные различия между наборами вопросов дерева решений для английского и русского языков. Таким образом, дерево решений для русского языка состоит из 38 общих вопросов плюс по одному вопросу для каждой единицы фонемного набора отдельно для левого и правого контекста.

3.2. Методика выбора оптимального фонемного набора. Размер используемого фонемного набора определяет количество контекстно-независимых моделей и также влияет на число контекстно-зависимых моделей. Если их число излишне велико, возрастает вычислительная сложность системы распознавания. Напротив, если их количество слишком мало, может снизиться точность системы, так как акустически схожие модели будут чаще распознаваться неправильно (спутываться). На данный момент существует два основных подхода к выбору такого набора для акустического моделирования: основанный на знаниях и статистический.

Таблица 2. Различия между наборами вопросов дерева решений для английского и русского языков

Удаленные вопросы	Добавленные вопросы для русского
Является ли левый/правый гласный звук долгим? Является ли левый/правый гласный звук коротким? Является ли левый/правый гласный звук дифтонгом? Является ли левый/правый гласный звук редуцированным?	Является ли левый/правый гласный звук ударным? Является ли левый/правый гласный звук безударным?
Является ли левый/правый согласный звук силлабическим? Является ли левый/правый согласный звук непрерывным?	Является ли левый/правый согласный звук мягким? Является ли левый/правый согласный звук твердым? Является ли левый/правый согласный звук дрожащим?

Для русского языка часто используются фонемные наборы, разработанные экспертами на основе лингвистических и фонологических правил [17], поскольку правила преобразования орфографического текста в фонемное представление для русского являются относительно несложными. Прямое преобразование графем в акустические единицы дает набор из 49 единиц, такой набор фонем использовался в [18] для сравнения с графемной системой распознавания речи. В работе [19] использовался набор из 43 фонологических единиц, который являлся стандартным алфавитом SAMPA для русского языка с добавлением фонемы /ɣ/. Для распознавания слитной русской речи с большим словарем в работе [20] был предложен фонемный набор из 59 единиц. В большинстве случаев исследователи используют расширенные наборы для представления гласных, включающие их ударные и безударные варианты [20-22].

Однако для других языков существуют исследования, в которых для получения фонемного набора используется статистическая информация. Для китайского языка было предложено использовать взаимную информацию между частями слов и их фонетическими транскрипциями в обучающем тексте [23]. Фонемные наборы создавались путем объединения тонально-зависимых фонем, которое приводит к минимальному уменьшению значения взаимной информации. Это позволило значительно сократить количество контекстно-зависимых моделей.

Для английского языка было предложено автоматически генерировать набор акустических единиц, основываясь на речевых данных [24].

Фонематические транскрипции в этом исследовании создавались с применением полученных акустических единиц, что позволило увеличить точность системы распознавания. Однако это является и главным недостатком системы из-за сложности добавления в нее новых слов.

В данной работе использовался комбинированный метод, в котором объединяется информация, полученная из фонологических знаний, и статистические данные от фонемной матрицы спутывания. Фонологические знания включают в себя правила произношения и информацию о фонологических чередованиях. Данный подход позволяет определить акустически близкие фонологические единицы. С помощью матрицы спутывания (см. рисунок 1) определяются наиболее часто несовпадающие монофоны. Для выбора наилучшего фонемного набора вначале используется наибольший набор, затем его размер постепенно уменьшается путем удаления или объединения некоторых фонологических единиц.

	a	a!	g	g'	e	e!	s	s'
a	6969	626	14	0	82	77	20	7
a!	685	3076	0	0	25	102	3	0
g	40	2	562	0	4	1	0	0
g'	2	0	1	38	1	0	1	0
e	121	11	5	0	1016	102	7	8
e!	100	40	1	0	125	2168	5	0
s	19	5	0	0	3	2	2248	59
s'	6	2	0	0	5	7	123	1293

Рис. 1. Пример матрицы спутывания монофонов (количество распознанных фонем внутри пары выделено квадратом)

Методика выбора фонем (см. рисунок 2) включает в себя следующие шаги:

1. Из алфавита IPA выбирается набор P0 путем применения фонологических правил произношения, чтобы найти наиболее акустически близкие единицы, например, гласные, различающиеся уровнем редукции.

2. В соответствии с фонологической спецификой языка определяются пары фонологических единиц, являющиеся кандидатами на объединение. Эти пары включают в себя мягкие/твердые согласные и ударные/безударные гласные.

3. С помощью набора P0 выполняется фонемное распознавание, и создается матрица спутывания монофонов. Для выбранных пар

вычисляется коэффициент спутывания (CR), который определяется следующим образом:

$$CR = \frac{M_1 + M_2}{H_1 + M_1 + H_2 + M_2} \cdot 100\%$$

где H_1 – количество правильно распознанных появлений первого монофона в паре (например, /a/ распозналась как /a/), H_2 – количество правильно распознанных появлений второго монофона в паре (например, /a!/ распозналась как /a!/), M_1 – количество неправильно распознанных появлений первого монофона в паре (например, /a/ распозналась как /a!/), M_2 – количество неправильно распознанных появлений второго монофона в паре (например, /a!/ распозналась как /a/). Чем больше коэффициент спутывания, тем больше монофоны в паре не совпадают, что делает их вероятными кандидатами для объединения.

4. Пары монофонов сортируются по уменьшению коэффициента спутывания.

5. Выбираются первые N пар монофонов, и после объединения соответствующих им фонологических единиц получается новый фонемный набор.

Выбор различных значений N позволяет получить несколько фонемных наборов. Наиболее подходящий набор можно определить, оценивая точность и скорость распознавания речи, полученную при его использовании.

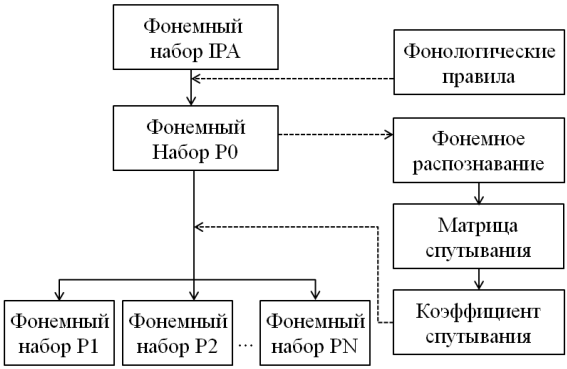

Рис. 2. Методика выбора фонем для системы автоматического распознавания русской речи

3.3. Моделирование произношения. Основными проблемами автоматического создания фонематических транскрипций для слов русского языка являются: плавающее ударение, замена буквы "ё" на

"е", большое количество омографов, а также явления редукции и ассимиляции звуков, возникающие в разговорной речи.

В данной работе фонематические транскрипции слов создавались автоматически путем применения к списку слов правил транскрибирования [25]. Для автоматической генерации транскрипций использовалась база данных словоформ русского языка с отмеченным ударением, которая была создана путем объединения двух баз данных, свободно доступных в Интернете: морфологической базы данных проекта STARLING (starling.rinet.ru) и морфологической базы данных проекта АОТ (www.aot.ru). Объем полученной базы данных составляет 2,3 млн словоформ. Подробнее процесс создания фонематических транскрипций слов описан в [26].

4. Базы данных слитной русской речи. Для обучения акустических моделей и проведения экспериментов по распознаванию русской речи использовались два речевых корпуса. Оба корпуса используют частоту дискретизации 16 кГц. Первый корпус – это корпус русской речи GlobalPhone [18], который содержит записи чтения русских газетных статей. Корпус состоит из 12321 фразы, в записи корпуса приняли участие 115 человек (61 мужчина и 54 женщины). Общий объем записей составил 26 часов. Из корпуса были удалены фразы, содержащие речевые сбои и шум, оставшаяся часть корпуса разделена на обучающий корпус (15 часов 25 минут) и тестовый корпус (1 час 40 минут), который содержит записи 5 мужчин и 5 женщин.

Второй корпус был записан в СПИИРАН [27]. В записи корпуса приняли участие 50 дикторов - носителей русского языка (25 женщин и 25 мужчин). Корпус содержит 327 фонетически сбалансированных фраз на русском языке (321 фраза состоит из одного предложения, остальные фразы являются короткими текстами, содержащими по несколько предложений). Общий объем корпуса -13,5 Гб, длительность записей – более 21 часа. Запись проводилась в специальной звукоизолированной комнате, отношение сигнал/шум составляло больше 35 дБ. Дополнительно для экспериментов по распознаванию русской речи были записаны 100 фраз, произнесенных двумя дикторами (мужчиной и женщиной). Длительность записи составила 30 минут. Фразы взяты из материалов интернет-газеты «Фонтанка.ru» (www.fontanka.ru). Каждая фраза в этом тестовом корпусе состоит из одного предложения длиной от 6 до 20 слов.

5. Описание экспериментов и их результаты.

5.1. Условия проведения эксперимента. Для каждого набора фонем была создана одна система распознавания речи, использующая монофоны, и три различные системы, использующие трифоны, в которых применялась одинаковая акустическая модель, но разные по раз-

меру словари и модели языка. В первой системе использовалась закрытая биграммная модель языка, обученная на транскрипциях речевой базы данных СПИИРАН, размер словаря составлял 1146 слов, значение коэффициента неопределенности было равно 241. Оценка системы производилась на тестовой части базы данных СПИИРАН. Вторая система построена аналогичным образом, но для обучения биграммной модели языка использовались транскрипции базы данных GlobalPhone, размер словаря составил 20 тыс. слов, значение коэффициента неопределенности – 137. Тестирование системы производилось на тестовой части базы данных GlobalPhone. Третья система использовала биграммную модель языка со словарем 204 тыс. слов, эта модель применялась для задачи распознавания речи со сверхбольшим словарем. Для оценки третьей системы также использовалась тестовая часть базы данных GlobalPhone, значение коэффициента неопределенности составило 844, количество внесловарных слов – 3,41%.

Акустические модели создавались с помощью инструментария HTK (Hidden Markov Model Toolkit) [15]. Для извлечения информативных признаков речевого сигнала использовались энергия сигнала и 12 мел-частотных кепстральных коэффициентов с их 1-й и 2-й производными. Данные признаки были вычислены путем анализа с помощью 26-канальной гребенки фильтров сегментов речи длительностью 20 мс с 10 мс перекрытием между сегментами. В качестве акустических моделей использовались лево-правые скрытые марковские модели с тремя состояниями.

5.2. Выбор фонемных наборов. На основе фонологических особенностей алфавита IPA был получен фонемный набор P0 путем выполнения следующих шагов [28]:

1. Ударные гласные звуки [a], [æ] и [ɑ] были объединены, поскольку они являются акустически близкими и различаются только от их местоположения в слове.

2. Безударные гласные звуки [ɐ] и [ə] были объединены, поскольку их произношения отличаются лишь незначительно в зависимости от расстояния от ударного слога. Аналогично были объединены [ʉ] и [ʊ].

3. Звук [ɵ] является похожим на комбинацию звуков [j] и [o], и соответственно он был представлен как 2 фонемы.

4. Согласные звуки [ʑ] и [ɣ] были исключены, поскольку они используются только в некоторых диалектах русского языка.

Получившийся фонемный набор P0 представлен в таблице 3. Символ «!» обозначает ударные гласные, символ «’» — мягкие (палатализованные) согласные.

Таблица 3. Фонемный набор P0

Классы фонологических единиц		Список фонологических единиц
Согласные	Твердые	/b/, /v/, /g/, /d/, /zh/, /z/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /t/, /f/, /h/, /c/, /sh/
	Мягкие	/b'/, /v'/, /g'/, /d'/, /z'/, /j/, /k'/, /l'/, /m'/, /n'/, /p'/, /r'/, /s'/, /t'/, /f'/, /h'/, /ch/, /sch/
Гласные	Ударные	/a!/, /e!/, /i!/, /o!/, /u!/, /y!/
	Безударные	/a/, /e/, /i/, /u/, /y/

Набор P0 содержит 47 фонологических единиц: 6 ударных и 5 безударных гласных и 36 согласных. В качестве кандидатов на объединение были выбраны пары твердых и мягких согласных, потому что они акустически схожи между собой, а также пары ударных и безударных гласных, поскольку они различаются только в длительности звука.

С использованием набора P0 была создана система распознавания фонем и выполнены эксперименты по их распознаванию. Из получившейся матрицы спутывания был сформирован список фонемных пар, отсортированный по величине коэффициента спутывания (см. рисунок 3). Из рисунка видно, что парами с наибольшим процентом спутывания являются пары гласных, поэтому они были объединены, в результате чего был получен фонемный набор P1.

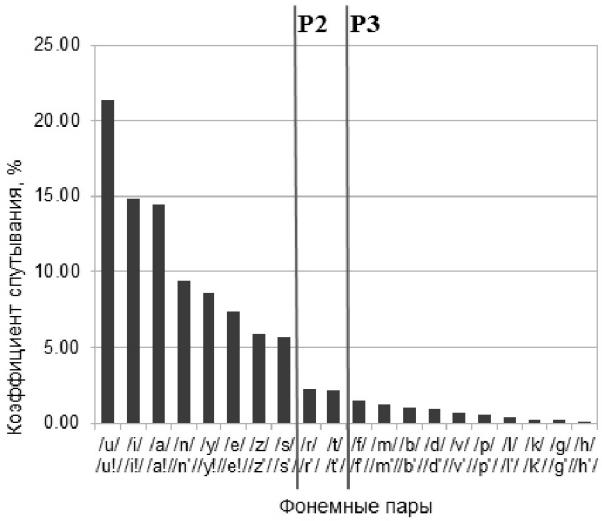

Рис. 3. Коэффициент спутывания для всех пар-кандидатов

Также из рисунка 3 видно, что пары согласных /n/-/n'/, /z/-/z'/, /s/-/s'/, /r/-/r'/ и /t/-/t'/ имеют более высокое значение процента спутывания, чем другие. Кроме того, различие в значении процента спутывания между парами /s/-/s'/ и /r/-/r'/ достаточно большое. Поэтому были сформированы три дополнительные фонемные наборы представлены в таблице 4.

Таблица 4. Описание созданных наборов фонем

Фонемный набор	Число фонологических единиц	Описание
P0	47	См. таблицу 3
P1	42	P0 без /a!/, /e!/, /i!/, /u!, /y!/
P2	39	P1 без /n'/, /s'/, /z'/
P3	37	P2 без /r'/, /t'/
P4	27	P3 без /b'/, /g'/, /d'/, /k'/, /l'/, /m'/, /p'/, /f'/, /h'/

Набор P2 сформирован путем объединения пар /n/-/n'/, /z/-/z'/, /r/-/r'/ и набор P3, дополнительно объединяющий пары /s/-/s'/ и /t/-/t'/. Также был создан набор P4, в котором были объединены все пары-кандидаты.

5.3. Оценивание эффективности применения фонемных наборов. Кроме системы распознавания, использующей фонемный набор P0, было создано еще 4 системы с фонемными наборами P1-P4. Точность распознавания фонем, полученная с использованием моделей монофонов, представлена в верхней части таблицы 5. В качестве модели языка использовалась фонемная биграммная модель, обученная на фонемных транскрипциях обеих баз данных. Тестовые части аудиокорпусов были также объединены в этом эксперименте. Наилучшая точность распознавания была достигнута при использовании набора P2. Большое различие в точности распознавания, полученной при использовании набора P0 и остальных наборов, объясняется отсутствием в наборах P1-P4 наиболее путающихся пар гласных. Низкая точность, полученная с набором P4, вызвана тем, что слишком много согласных было объединено, что привело к уменьшению разрешения пространства фонологических единиц. Результаты экспериментов показали, что пары-кандидаты были выбраны правильно, и их объединение дает наилучшие результаты.

В нижней части таблицы 5 приведены результаты экспериментов, полученные с применением трифонных моделей. Биграммная модель языка была обучена на транскрипциях как обучающих, так и тестовых частей речевых корпусов, таким образом, в данном случае используется модель языка с закрытым словарем. Однако обучающие и тестовые части корпуса акустически различны. Результаты распозна-

вания, полученные с применением различных фонемных наборов, приблизительно одинаковые, но при использовании набора P0 точность оказалась чуть выше. Высокая точность распознавания, полученная для базы данных СПИИРАН, связана с тем, что все дикторы произносили одни и те же фразы, то есть лексическое содержание фраз было одинаковым, а также размер словаря значительно меньше, чем в речевом корпусе GlobalPhone.

Таблица 5. Точность распознавания речи, полученная при использовании различных фонемных наборов

Тестовые корпуса/ Фонемные наборы	P0	P1	P2	P3	P4
Точность распознавания фонем, %					
СПИИРАН + GlobalPhone	48,92	52,43	53,20	53,06	52,93
Точность распознавания слов, %					
СПИИРАН	96,64	96,55	96,41	96,62	96,40
GlobalPhone	80,92	80,35	80,67	79,82	78,52

Количество контекстно-зависимых трифонов для каждого фонемного набора представлено в таблице 6, где также указано количество связанных состояний соответствующей акустической модели.

Таблица 6. Количество трифонов и состояний для различных фонемных наборов

Набор фонем	Количество трифонов	Число состояний
P0	112849	5342
P1	81314	5356
P2	65562	5359
P3	56279	5335
P4	22709	5342

Поскольку различие в точности распознавания слов, полученной при применении различных фонемных наборов, для речевой базы данных СПИИРАН оказалось достаточно небольшим (менее 1% в абсолютных значениях), были проведены дополнительные эксперименты по определению скорости распознавания речи. В качестве показателя скорости распознавания использовался показатель реального времени (real time factor - RTF). На рисунке 4 представлен график зависимости точности распознавания слов от значения RTF. Для значения RTF=1,2 и ниже фонемный набор P3 показал лучшие результаты распознавания, в то время как для больших значений RTF (более медленное распознавание) лучшие результаты показал набор P0.

Высокая точность распознавания слов объясняется тем, что применялась закрытая модель языка с малым размером словаря. Для создания открытой модели языка со сверхбольшим размером словаря использовалась статистическая биграммная модель, описанная в работе [29]. Результаты, полученные на тестовой части корпуса GlobalPhone, представлены в таблице 7. Хотя точность распознавания существенно ниже, полученные результаты показывают такую же тенденцию, как и предыдущие эксперименты.

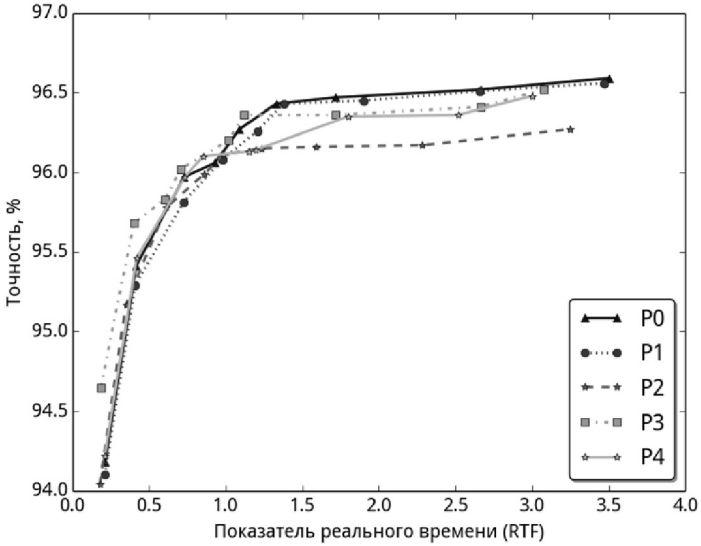

Рис. 4. Точность распознавания слов в зависимости от значения RTF, полученная с использованием речевого корпуса СПИИРАН

Таблица 7. Точность распознавании речи, полученная при использовании модели языка с открытым словарем

Фонемный набор	P0	P1	P2	P3	P4
	67,07	66,78	66,53	65,93	65,28

Таким образом для дальнейших экспериментов по распознаванию речи со сверхбольшим словарем с трифонными акустическими моделями был выбран набор P0, состоящий из 47 фонологических единиц.

5.4. Результаты экспериментов по распознаванию слитной русской речи со сверхбольшим словарем. Во всех предыдущих экспериментах использовались стандартные биграмные языковые модели.

Нами была также создана синтактико-статистическая модель языка на основе базовой биграммной модели, полученной в результате автоматического статистического анализа обучающего текстового корпуса, собранного с интернет-сайтов электронных газет. Базовая биграммная модель была расширена за счет выполнения синтаксического анализа обучающего текстового корпуса, в ходе которого выявляются грамматически связанные пары слов, разделенные в тексте другими словами. Таким образом, синтаксический анализ позволяет учесть дальнодействующие грамматические связи между словами. Затем была проведена линейная интерполяция статистической триграммной модели с синтаксическо-статистической моделью с коэффициентом интерполяции 0,27. Размер словаря составил 204 тыс. слов, относительное количество внесловарных слов для тестовых данных составило 2,5 %. При создании модели языка использовалось сглаживание Несер-Нея (Kneser-Ney). Подробно процесс создания синтаксическо-статистической модели описан в [29, 30].

Для экспериментов с модифицированной моделью языка использовался декодер речи Julius 4.2 [31]. Для настроек параметров системы были использованы записи 4 дикторов из тестовой части GlobalPhone общей длительностью 40 минут. Записи остальных дикторов были объеденены с дополнительной частью базы данных СПИИРАН в тестовый набор, включающий в себя записи 8 дикторов (4 мужчин и 4 женщин) общей длительностью 1,5 часа.

Результаты распознавания слов и графем (под графемами понимаются буквы и знак пробела) представлены в таблице 8, в которой также даны значения коэффициента неопределенности, относительного количества совпадений *n*-грамм (т.е. количества *n*-грамм в тестовых данных, которые присутствуют в модели языка) и относительного количества внесловарных слов для тестовых данных.

Таблица 8. Результаты распознавания слитной русской речи со сверхбольшим словарем

Коэффициент неопределенности модели языка	Относительное кол-во совпадений n-грамм, %	Точность распознавания слов, %	Точность распознавания графем, %
549	39,2 (+44,0 для 2-грамм)	73,1	91,5

Скорость распознавания речи на компьютере с многоядерным процессором с тактовой частотой 1,8 ГГц составила около 2,0 RTF.

6. Заключение. В статье представлен выбор оптимального фонемного набора для системы распознавания слитной русской речи. Для процесса акустического моделирования было разработано новое фоне-

тическое дерево решений для создания трифонных моделей, учитывающее специфику фонологии русского языка. Была предложена комбинированная методика выбора оптимального фонемного набора, в которой используются как фонетические знания, так и статистическая информация. Эксперименты показали, что разница в точности распознавания слов системами с разными фонемными наборами незначительна, кроме системы с набором P4, в котором все пары-кандидаты были объединены. В оффлайн системах распознавания речи, где точность является важнейшим фактором, может быть рекомендовано использование фонемного набора P0, поскольку он позволяет достичь наибольшей точности. В системах распознавания реального времени, в которых необходим компромисс между скоростью и точностью распознавания, может быть рекомендовано применение набора P3, поскольку точность распознавания соответствующей системы при RTF $\approx$ 1 была выше.

В ходе экспериментов по распознаванию речи с применением синтактико-статистической языковой модели со сверхбольшим словарем, где использовался фонемный набор P0, состоящий из 47 фонем, наибольшая точность распознавания слов русской речи составила 73,1%. Полученные результаты соответствуют качеству распознавания слитной русской речи, полученному на настоящий момент в мире другими организациями.

Литература

1. *Whittaker E.W.D.* Statistical language modelling for automatic speech recognition of Russian and English // Ph.D. thesis. Cambridge Univ. 2000. 140 p.
2. *Karpov A., Kipyatkova I., Ronzhin A.* Speech Recognition for East Slavic Languages: The Case of Russian // Proceedings of 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages SLTU'2012. Cape Town. RSA. 2012. pp. 84–89.
3. *Кипяткова И.С., Карпов А.А.* Аналитический обзор систем распознавания русской речи с большим словарем // Труды СПИИРАН. 2010. №1(12). С.7–20.
4. *Kipyatkova I., Karpov A.* Study of Morphological Factors of Factored Language Models for Russian ASR / Edited by A. Ronzhin et al. // SPECOM 2014. Springer International Publishing Switzerland. LNAI 8773. 2014. pp. 451–458.
5. *Vazhenina D., Kipyatkova I., Markov K., Karpov A.* State-of-the-art Speech Recognition Technologies for Russian Language // Proceedings of the Joint International Conference on Human-Centered Computer Environments HCCE-2012. Aizu-Wakamatsu. Japan. 2012. pp. 59–63.
6. *Titov Y., Kilgour K., Stüker S., Waibel A.* The 2011 kit quaero speech-to-text system for the Russian language // Proceedings of the 14th International Conference “Speech and Computer” (SPECOM’2011). 2011. pp. 136–143.
7. *Shin E., Stüker S., Kilgour K., Fügen C., Waibel A.* Maximum Entropy Language Modeling for Russian ASR // Proceedings of the International Workshop for Spoken Language Translation (IWSLT 2013). 2013.

8. *Zablotskiy S., Shvets A., Sidorov M., Semenkin E., Minker W.* Speech and Language Recources for LVCSR of Russia // Proceedings of LREC'2012. Istanbul. Turkey. 2012. pp. 3374–3377.
9. *Zulkarneev M., Satunovsky P., Shamraev N.* The use of d-gram language model for speech recognition in Russian // SPECOM 2013. Springer LNAI 8113. 2013. pp. 362–366.
10. *Zulkarneev M., Grigoryan R., Shamraev N.* Acoustic modeling with deep belief networks for Russian Speech // SPECOM 2013. Springer LNAI 8113. 2013. pp. 17–23.
11. *Schalkwyk J., Beeferman D., Beaufays F., Byrne B., Chelba C., Cohen M., Kamvar M., Strope B.* Google Search by Voice: A Case Study // Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics. 2010. pp. 61–90.
12. SpeechKit API. URL: http://api.yandex.ru/speechkit/ (дата обращения: 25.04.2014).
13. *Levin K., Ponomareva I., Bulusheva A., Chernykh G., Medennikov I., Merkin N., Prudnikov A., Tomashenko N.* Automated closed captioning for Russian live broadcasting // Proceedings of Interspeech'2014. 2014. pp. 1438–1442.
14. *Young S., Odell J., Woodland P.* Tree-based state tying for high accuracy acoustic modeling // Proceedings of Int. Workshop on Human Language Technology HLT'1994. Stroudsburg. PA. USA. 1994. pp. 307–312.
15. *Young S. et al.* The HTK book // Cambridge Univ. Press. 2009. 384 p.
16. *Odell J.* The use of context in large vocabulary speech recognition // Ph.D. thesis. Cambridge Univ., 1995.
17. *Cubberley P.* Russian: a linguistic introduction // Cambridge University Press. 2002.
18. *Schultz T., Waibel A.* Development of Multilingual acoustic models in the GlobalPhone project // Proceedings of TSD'1998. Brno. Czech Republic. 1998. pp. 311–316.
19. *Psutka J., Ircing P., Psutka J.V., Hajic J., Byrne W.J., Mirovsky J.* Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project // Proceedings of Interspeech'2005. Lisbon. Portugal. 2005. pp. 1349–1352.
20. *Tatarnikova M., Tampel I., Oparin I., Khokhlov Y.* Building acoustic models for a large vocabulary continuous speech recognizer for Russian // Proceedings of SPECOM'2006. St. Petersburg, Russia. 2006. pp. 83–87.
21. *Stüker S., Schultz T.* A grapheme based speech recognition system for Russian // Proceedings of International Conference SPECOM'2004. St. Petersburg, Russia. 2004. pp. 297–303.
22. *Ronzhin A., Karpov A.* Russian Voice Interface // Pattern Recognition and Image Analysis. 2007. vol. 17. no. 2. pp. 321–336.
23. *Zhang J.S., Hu X.H., Nakamura S.* Using mutual information criterion to design an efficient phoneme set for Chinese speech recognition // IEICE Transactions on Information and Systems. 2008. vol. E91-D. no. 3. 2008. pp. 508–513.
24. *Singh R., Raj B., Stern R.* Automatic generation of subword units for speech recognition systems // IEEE Transactions on Acoustics. Speech and Signal Processing. 2002. vol. 10(2). pp. 89–99.
25. *Шведова Н.Ю. (гл. ред.) и др.* Русская грамматика // М.: Наука. 1980. 783 с.
26. *Кипяткова И.С., Карпов А.А.* Модуль фонематического транскрибирования для системы распознавания разговорной русской речи // Искусственный интеллект, Донецк. Украина. 2008. № 4. 2008. С. 747–757.
27. *Jokisch O., Wagner A., Sabo R., Jaeckel R., Cylwik N., Rusko M., Ronzhin A., Hoffmann R.* Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system // Proceedings of SPECOM'2009. St. Petersburg. Russia. 2009. pp. 515–520.
28. *Vazhenina D., Markov K.* Phoneme set selection for Russian speech recognition // Proceedings of Int. Conf. on Natural Language Processing and Knowledge Engineering NLP-KE. Tokushima. Japan. 2011. pp. 475–478.

29. *Karpov A., Markov K., Kipyatkova I., Vazhenina D., Ronzhin A.* Large vocabulary Russian speech recognition using syntactico-statistical language modeling // Speech Communication. 2014. vol. 56. pp. 213–228.
30. *Кипяткова И.С.* Программно-алгоритмическое обеспечение создания синтаксическо-статистической модели русского языка по текстовому корпусу // Труды СПИИРАН. 2013. Вып. 24. С. 332–348.
31. *Lee A., Kawahara T.* Recent Development of Open-Source Speech Recognition Engine Julius // Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2009). Sapporo. Japan. 2009. pp.131–137.

References

1. Whittaker E.W.D. Statistical language modelling for automatic speech recognition of Russian and English. Ph.D. thesis. Cambridge Univ. 2000. 140 p.
2. Karpov A., Kipyatkova I., Ronzhin A. Speech Recognition for East Slavic Languages: The Case of Russian. Proceedings of 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages SLTU'2012. Cape Town. RSA. 2012. pp. 84–89.
3. Kipyatkova I.S., Karpov A.A. [An Analytical Survey of Large Vocabulary Russian Speech Recognition Systems]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2010. vol. 12. pp.7–20. (In Russ.).
4. Kipyatkova I., Karpov A. Study of Morphological Factors of Factored Language Models for Russian ASR. Edited by A. Ronzhin et al. SPECOM 2014. Springer International Publishing Switzerland. LNAI 8773. 2014. pp. 451–458.
5. Vazhenina D., Kipyatkova I., Markov K., Karpov A. State-of-the-art Speech Recognition Technologies for Russian Language. Proceedings of the Joint International Conference on Human-Centered Computer Environments (HCCE-2012). Aizu-Wakamatsu. Japan. 2012. pp. 59–63.
6. Titov Y., Kilgour K., Stüker S., Waibel A. The 2011 kit quaero speech-to-text system for the Russian language. Proceedings of the 14th International Conference "Speech and Computer" (SPECOM'2011). 2011. pp. 136–143.
7. Shin E., Stüker S., Kilgour K., Fügen C., Waibel A. Maximum Entropy Language Modeling for Russian ASR. Proceedings of the International Workshop for Spoken Language Translation (IWSLT 2013). 2013.
8. Zablotskiy S., Shvets A., Sidorov M., Semenkin E., Minker W. Speech and Language Recources for LVCSR of Russia. Proceedings of LREC'2012. Istanbul. Turkey. 2012. pp. 3374–3377.
9. Zulkarneev M., Satunovsky P., Shamraev N. The use of d-gram language model for speech recognition in Russian. SPECOM 2013. Springer LNAI 8113. 2013. pp. 362–366.
10. Zulkarneev M., Grigoryan R., Shamraev N. Acoustic modeling with deep belief networks for Russian Speech. SPECOM 2013. Springer LNAI 8113. 2013. pp. 17–23.
11. Schalkwyk J., Beeferman D., Beaufays F., Byrne B., Chelba C., Cohen M., Kamvar M., Strope B. Google Search by Voice: A Case Study. Advances in Speech Recognition: Mobile Environments. Call Centers and Clinics. 2010. pp. 61–90.
12. SpeechKit API. Available at: http://api.yandex.ru/speechkit/ (accessed 25.04.2014).
13. Levin K., Ponomareva I., Bulusheva A., Chernykh G., Medennikov I., Merkin N., Prudnikov A., Tomashenko N. Automated closed captioning for Russian live broadcasting. Proceedings of Interspeech'2014. 2014. pp. 1438–1442.
14. Young S., Odell J., Woodland P. Tree-based state tying for high accuracy acoustic modelling. Proceedings of Int. Workshop on Human Language Technology (HLT'1994). Stroudsburg. PA. USA. 1994. pp. 307–312.
15. Young S. et al. The HTK book. Cambridge Univ. Press. 2009. 384 p.

16. Odell J. The use of context in large vocabulary speech recognition. Ph.D. thesis. Cambridge Univ. 1995.
17. Cubberley P. Russian: a linguistic introduction. Cambridge University Press. 2002.
18. Schultz T., Waibel A. Development of Multilingual acoustic models in the GlobalPhone project. Proceedings of TSD'1998. Brno. Czech Republic. 1998. pp. 311–316.
19. Psutka J., Ircing P., Psutka J.V., Hajic J., Byrne W.J., Mirovsky J. Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project. Proceedings of Interspeech'2005. Lisbon. Portugal. 2005. pp. 1349–1352.
20. Tatarnikova M., Tampel I., Oparin I., Khokhlov Y. Building acoustic models for a large vocabulary continuous speech recognizer for Russian. Proceedings of SPECOM'2006. St. Petersburg. Russia. 2006. pp. 83–87.
21. Stüker S., Schultz T. A grapheme based speech recognition system for Russian. Procceedings of International Conference SPECOM'2004. St. Petersburg. Russia. 2004. pp. 297–303.
22. Ronzhin A., Karpov A. Russian Voice Interface. Pattern Recognition and Image Analysis. 2007. vol. 17. no. 2. pp. 321–336.
23. Zhang J.S., Hu X.H., Nakamura S. Using mutual information criterion to design an efficient phoneme set for Chinese speech recognition. IEICE Transactions on Information and Systems. 2008. vol. E91-D. no. 3. pp. 508–513.
24. Singh R., Raj B., Stern R. Automatic generation of subword units for speech recognition systems. IEEE Transactions on Acoustics, Speech and Signal Processing. 2002. vol. 10(2). pp. 89–99.
25. Shvedova N.Yu. et. al. Russkaya grammatika [Russian Grammar]. M.: Nauka, 1980. 783 p. (In Russ.)
26. Kipyatkova I.S., Karpov A.A. [The module of phonemic transcription for conversational Russian speech recognition system]. *Iskusstvenny intellekt – Artificial intelligence.* Donetsk. Ukraine. 2008. vol. 4,. pp. 747–757. (In Russ.).
27. Jokisch O., Wagner A., Sabo R., Jaeckel R., Cylwik N., Rusko M., Ronzhin A., Hoffmann R. Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system. Proceedings of SPECOM'2009. St. Petersburg. Russia. 2009. pp. 515–520.
28. Vazhenina D., Markov K. Phoneme set selection for Russian speech recognition. Proceedings of Int. Conf. on Natural Language Processing and Knowledge Engineering NLP-KE. Tokushima. Japan. 2011. pp. 475–478.
29. Karpov A., Markov K., Kipyatkova I., Vazhenina D., Ronzhin A. Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Communication. 2014. vol. 56. pp. 213–228.
30. Kipyatkova I.S. [Software for Creation of Sintactico-Statistical Russian Language Model Based on the Text Corpus]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2013. vol. 24. pp.332–348. (In Russ.).
31. Lee A., Kawahara T. Recent Development of Open-Source Speech Recognition Engine Julius. Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2009). Sapporo. Japan. 2009. pp.131–137.

Важенина Дарья Анатольевна — магистр техн. наук, аспирант лаборатории пользовательских интерфейсов отделения информационных систем Университета Аизу, Япония. Область научных интересов: автоматическое распознавание речи, статистические языковые модели, машинное обучение. Число научных публикаций — 8. d8132102@u-aizu.ac.jp. Университет Аизу, Тсуруга, Икки-мачи, г.Аизу-Вакаматсу, преф.Фукушима, 965-8580, Япония; р.т.+81(242)372762, факс: +81(242)372549.

Vazhenina Daria Anatolievna — M.Sci., Ph.D. student, Human Interface Laboratory of the Information Systems Division, University of Aizu, Japan. Research interests: automatic speech recognition, statistical language modeling, machine learning. The number of publications — 8. d8132102@u-aizu.ac.jp; The University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu-shi, Fukushima-ken, 965-8580, Japan; fax: +81(242)372549.

Кипяткова Ирина Сергеевна — к-т техн. наук, старший научный сотрудник лаборатории речевых и многомодальных интерфейсов Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: автоматическое распознавание речи, статистические модели языка. Число научных публикаций — 50. kipyatkova@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Kipyatkova Irina Sergeevna — Ph.D., senior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition, statistical language models. The number of publications — 50. kipyatkova@iias.spb.su; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Марков Константин — к-т техн. наук, доцент лаборатории пользовательских интерфейсов отделения информационных систем Университета Аизу, Япония. Область научных интересов: обработка аудио сигналов, Байесовские статистические модели, машинное обучение, распознавание образов. Число научных публикаций — 100. markov@u-aizu.ac.jp; Университет Аизу, Тсуруга, Икки-мачи, г.Аизу-Вакаматсу, преф.Фукушима, 965-8580, Япония; р.т.+81(242)372762, факс: +81(242)372549.

Markov Konstantin — Ph.D., associate professor, Human Interface Laboratory of the Information Systems Division, University of Aizu, Japan. Research interests: audio signal processing, Bayesian statistical modeling, machine learning and pattern recognition. The number of publications — 100. markov@u-aizu.ac.jp; The University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu-shi, Fukushima-ken, 965-8580, Japan; office phone +81(242)372762, fax: +81(242)372549.

Карпов Алексей Анатольевич — д-р техн. наук, доцент, ведущий научный сотрудник лаборатории речевых и многомодальных интерфейсов Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: автоматическое распознавание речи, многомодальные интерфейсы, аудиовизуальное распознавание речи. Число научных публикаций — 190. karpov@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Karpov Alexey Anatolyevich — Ph.D., Dr. Sci., associate professor, leading researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition, multimodal interfaces, audio-visual speech recognition. The number of publications — 190. karpov@iias.spb.su; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Поддержка исследований. Работа выполнена при финансовой поддержке фонда РФФИ (проект № 12-08-01265-а).

Acknowledgements. This research is supported by RFBR (project № 12-08-01265-a).

РЕФЕРАТ

Важенина Д.А., Кипяткова И.С., Марков К., Карпов А.А. **Методика выбора фонемного набора для автоматического распознавания русской речи.**

Статья посвящена обучению акустических моделей и выбору оптимального фонемного алфавита. Для создания моделей акустических единиц было разработано фонетическое дерево решений, учитывающее специфику фонологии русского языка. Для выбора оптимального фонемного набора предложен комбинированный метод, который объединяет информацию, полученную из фонологических знаний, и статистические данные из матрицы спутывания монофонов. Матрица спутывания позволяет определить наиболее часто несовпадающие монофоны. Для выбора наилучшего фонемного набора вначале использовался наибольший фонемный набор, затем его размер постепенно уменьшался путем удаления или объединения некоторых фонологический единиц. Применяя данный метод к русскому варианту IPA/МФА, фонемный набор был уменьшен до 47 фонологических единиц и был использован как исходный. За счет применения матрицы спутывания были созданы 4 набора с разным количеством фонологических единиц: 42, 39, 37, 27. Система распознавания речи была обучена с использованием каждого фонемного набора. Эксперименты по распознаванию русской речи, которые проводились на речевых корпусах СПИИРАН и GlobalPhone, показали, что использование сокращенных фонемных наборов позволяет увеличить точность распознавания фонем и практически не влияет на точность распознавания слов. При этом использование сокращенного фонемного набора позволяет повысить скорость распознавания речи, что важно для систем, работающих в реальном времени.

Для проведения экспериментов по распознаванию русской речи со сверхбольшим словарем был выбран набор, состоящий из 47 фонологических единиц. Распознавание речи проводилось с использованием синтаксическо-статистической модели языка, которая была создана путем добавления к базовой биграммной модели синтаксически связанных пар слов, которые были разделены в обучающем тексте другими словами. Размер словаря составил 204 тыс. слов. В результате экспериментов точность распознавания слов составила 73,1%.

SUMMARY

Vazhenina D.A., Kipyatkova I.S., Markov K., Karpov A.A. **Technique for phoneme set selection for automatic Russian speech recognition.**

The paper is devoted to training the acoustic models and selection of the optimal phoneme alphabet. The phonetic tree that takes into account specifics of the Russian phonology was developed to create context-dependent acoustic models. A combined method, which utilizes information obtained from phonological knowledge and statistical data from the confusion matrix of monophones, was proposed for selection of the optimal phoneme set. Confusion matrix allows to determine the most frequently mismatched monophones. For the best phoneme set selection, the largest phoneme set was used and then its size was gradually decreases by deleting or merging some phonemes. Applying this approach to the IPA Russian phonetic set, we first reduced it to 47 phonological units, which were used as initial set. Based on the phoneme confusion results 4 sets with the number of phonological units of 42, 39, 27, 27 were created. The speech recognition systems were trained separately using each phoneme set. Experiments on Russian speech recognition conducted on SPIIRAS and GlobalPhone speech corpora showed that usage of reduced phoneme sets allows to increase phoneme recognition accuracy and almost has no influence on word recognition accuracy. At the same time using the reduced phoneme set allows to increase decoding speed what is important for real-time systems.

For experiments concerning very large vocabulary Russian speech recognition the set consisting of 47 phonological units was selected. Speech recognition was conducted using the syntactico-statistical language model. This model was created by adding grammatically-connected word pairs, which were separated by other words in the training corpus, to the baseline bigram model. Vocabulary size was 204 K words. Word recognition accuracy of 73.1 % was achieved.