Моделирование вариативности произношения для уменьшения уровня ошибок при распознавании речи*

***В. Я. Чучупал*[1], *А. А. Коренчиков*[2]**
chuchu@ccas.ru
[1]Москва, Вычислительный Центр им. А. А. Дородницына РАН; [2]Московский государственный университет им. М. В. Ломоносова

Рассматривается возможность снижения уровня ошибок при автоматическом распознавании русской речи за счет использования моделей вариативности произношения. Определена вероятностная модель вариативности произношения, способы оценки ее параметров и реализации в рамках стандартных процедур распознавания речи. Показано, что использование явных моделей вариативности произношения может быть эффективным способом снижения уровня ошибок при распознавании русской разговорной речи, в том числе при несоответствии характеристик обучающего и тестового речевого материала.

Ключевые слова: распознавание речи; акустическое моделирование; вариативность речи; моделирование произношения; скрытые марковские модели

Improving speech recognition accuracy by means of word pronunciation modeling*

***V. J. Chuchupal*[1], *A. A. Korenchikov*[2]**
[1]Dorodnicyn Computing Centre of Russian Academy of Sciences, Moscow; [2]Lomonosov Moscow State University, Moscow

Background: Pronunciation variation modeling evidently has a big potential as a simple way to significantly improve the accuracy of automatic speech recognition. At the same time, the reported improvements in accuracy obtained with pronunciation variation models in experiments are still far from the expected ones.
Methods: The advantages of the so-called explicit pronunciation variation models are explored as an approach for improvement of natural Russian speech recognition accuracy. The probabilistic pronunciation variation model is formally defined as well as the methods of its parameter estimation.
Results: The effect of use of explicit pronunciation variation models is shown to be very dependent on the speech material type. Evaluation on the corpus with Russian read and planned speech shows a negligible effect of using the models. At the same time, the evaluation of pronunciation models on spontaneous Russian speech reveals substantial improvement of automatic speech recognition accuracy.
Conclusions: Despite big promises, there are a lot of efforts necessary to develop pronunciation variation models for speech recognition that will effectively account for speaker and speaking style, accents, and dialects. Nevertheless, right now, the pronunciation model of explicit type can show substantial improvement of recognition accuracy on natural speech recognition task.

Keywords: automatic speech recognition; acoustic modeling; speech variativity; pronunciation modeling; hidden markov models

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-01-00607

Введение

Произнесение слова в системах распознавания слитной речи определяется заданием его произносительной транскрипции: последовательности составляющих это слово фонем. Большинство слов в словаре систем распознавания речи имеет один вариант произношения — каноническую или базовую транскрипцию, которая соответствует нормативному произношению. В повседневной разговорной речи произношение слов может отличаться от нормативного, что является одной из основных причин ошибок при автоматическом распознавании.

Под моделированием вариативности произношения в речевой технологии подразумевают разработку методов определения множества наиболее вероятных акустических образов слов и их последовательностей.

В литературе встречаются два основных подхода к моделированию вариативности произношения [1, 2]. Явное моделирование (explicit modeling) заключается в моделировании вариативности произнесения путем описания возможных изменений в фонемной транскрипции слов [2]. Модель вариативности произношения в данном случае связана с множеством произносительных транскрипций слова. Неявное моделирование (implicit modeling) [3] описывает вариативность произнесения путем изменений в структуре моделей звуков в канонической транскрипции слов, т. е. фонемная транскрипция у слова может быть одна, но имеет сложную форму, например, в виде графа из фонем.

Оба этих подхода не отменяют использования базовых, канонических транскрипций и направлены на определение дополнительных вариантов произнесения слов и словосочетаний.

Использование моделей вариативности произношения имеет высокий потенциал как способ повышения эффективности автоматического распознавания речи. Это очевидно при эвристическом анализе причин появления ошибок и подтверждается данными так называемых модельных экспериментов, когда за счет использования определенных экспертным образом произносительных транскрипций уровень пословной ошибки распознавания — WER (word error rate [4]) может быть уменьшен почти вдвое [5].

Фактические результаты применения моделей вариативности не соответствует ожиданиям. Так в работе [2] на материале корпуса VIOS (для голландского языка) уровень ошибки WER в лучшем случае снизился на 0,8% (с исходных 10,7% до 9,9%), при значительном (4,9 на слово, в среднем) числе допустимых вариантов произнесения. В экспериментах на корпусе данных Switchboard [3] за счет использования неявных моделей вариативности показатель WER снизился на 1,7% (с 39,4% до 37,7%). На корпусе данных NIST 2000 Hub-5 использование моделей вариативности для учета темпа речи дало уменьшение показателя WER на 2,2%: с 54,6% до 52,4% [6].

В приведенных выше работах моделирование, в частности, оценка параметров акустических и языковых моделей, основано на байесовском подходе и использовании моделей порождающего типа. Можно ожидать улучшения результатов при применении дискриминантных методов, которые до недавнего времени в такой задаче не использовались из–за отсутствия достаточных по размеру выборок данных. К настоящему времени для ряда языков уже есть достаточные по объему корпуса данных, собранные, например, в Google и Microsoft.

В работе [7] предложена и исследована дискриминантная модель вариативности произношения при наличии диалекта, в данном случае африканского английского. Вначале экспертным методом были найдены контекстные правила генерации возможных вариантов произнесения по базовым транскрипциям, а затем с использованием т.н. транскрип-

ций от оракула (наиболее корректные из списка N лучших при распознавании), обучалась нейронная сеть для присваивания весов правилам, что позволяло затем вводить веса для получаемых на основе этих правил произносительных вариантов слов.

Поскольку наличие диалекта, как правило, коррелирует с синтаксическими отклонениями, в данной работе одновременно с фонемными транскрипциями проводилась адаптация и языковых моделей.

Теоретически возможное улучшение показателя WER (для транскрипций от оракула) в данном эксперименте было более 10% : с 38,2% до 28,1%. Фактически же внедрение моделей произношения совместно с адаптацией модели языка дало улучшение WER на 2,1%, при этом основной вклад в это улучшение внесла адаптация языковых моделей, модели вариативности произношения принесли 0,6%. Аналогичная модель для оптимизации транскрипций американского стандартного языка также позволила уменьшить WER на 0,8% (с 25,2% до 24,5%).

Наличие огромных выборок данных дает возможность оценить эмпирические частоты транскрипций, что в принципе является оптимальным вариантом при использовании явных методов учета вариативности. В работе [8] на корпусе данных «живого» поиска Windows на мобильных устройствах (Windows Live Search for Mobile voice search task) предложено и экспериментально исследовано несколько способов моделирования вариативности произношения. Рассмотрена эмпирическая модель, когда вероятность фонемных транскрипций слова оценивалась через их относительные частоты, а также нескольких вариантов параметрических моделей со сглаживанием - для возможности оценки вероятности не встречавшихся в обучающей выборке вариантов. Сами варианты произнесения генерировались данным с помощью пофонемного распознавания. В результате удалось (для лучшей из предложенных, линейной комбинации эмпирической и параметрической, моделей) уменьшить величину показателя WER с 34,8% до 33,0%, т. е. на 1,8% или на 5,2%, если рассматривать относительное изменение ошибки.

Приведенные значения показателей эффективности распознавания показывают, что фактические результаты применения моделей вариативности приводят в получению весьма далеких от теоретически ожидаемых результатов, эта ситуация существенно не меняется последние десятилетия [9].

Данная работа во многом вызвана личном опытом авторов. При обработке русской разговорной речи, например выделении ключевых слов, результаты можно заметно улучшить за счет добавления произносительных вариантов для проблемных слов. В представленной работе исследована возможность снижения уровня ошибок автоматического распознавания русской речи за счет использования моделей вариативности произношения. Нас интересовало, насколько можно улучшить результаты распознавания за счет использования явного подхода к моделированию вариативности произношения и в каких случаях. Явный подход был выбран поскольку в этом случае изменения в существующих процедурах распознавания минимальны.

В данном случае мы следовали явному подходу к моделированию вариативности произношения, т. е. выбору наиболее вероятных транскрипций, предполагая, что все изменения в фактическом произнесении можно адекватно описать соответствующими фонемными транскрипциями. Поскольку технологии описания произношения в системах распознавания речи основана на использовании транскрипций, то появление ошибки означает, что фонемные транскрипции корректных слов оказались менее правдоподобными, чем фонемные транскрипции каких-то других слов словаря.

Далее в тексте термины модель произношения слова и его фонемная транскрипция — синонимы, модель вариативности произношения в таком случае соответствует некоторому множеству фонемных транскрипций.

Реализация явного подхода для моделирования вариативности произнесения слов в системе распознавания речи связана с решением следующих задач:

- определение вероятных вариантов произнесения слов словаря,
- оценка параметров модели вариативности,
- результативное использование вариантов произнесения при распознавании.

Модель вариативности произношения

Цель использования модели вариативности произношения в системе распознавания речи — уменьшение числа ошибок распознавания. В данном случае это предполагается достичь за счет нахождения и использования транскрипций, которые больше соответствуют фактически вариантам произнесения, чем базовые.

Различие между словами и транскрипциями заключается в том, что слова относятся к смыслу высказывания, а их произносительные транскрипции определяют акустические параметры и образы слов.

Это различие можно учесть путем детализации формулы классического вероятностного подхода к распознаванию речи [10].

Пусть $X = \{x_t\}, t = 1, \ldots, T$ — наблюдаемый образ в виде последовательности параметров речевого сигнала, а $W = \{w_i\}, i = 1, \ldots, N$ – последовательность слов словаря. Результат распознавания образа X, наиболее вероятную последовательность произнесенных слов W^*, можно определить из уравнения [11]

$$W^* = \arg\max_W P(W|X) = \arg\max_W \frac{P(X|W)P(W)}{P(X)}. \quad (1)$$

Первый сомножитель $P(X|W)$ в числителе (1) соответствует правдоподобию данных при заданной последовательности слов и определяется с помощью акустических моделей. Полученная величина правдоподобия затем умножается на значение $P(W)$, которое определяется с помощью модели языка. Знаменатель $P(X)$ — вероятность наблюдения X, выполняет функции нормализующего члена.

Пусть акустической моделью произнесения некоторого слова w служит его фонемная транскрипция t^w. Множество всех транскрипций слова w обозначим T^w. Моделью произнесения последовательности слов W будет любая последовательность их транскрипций, обозначим все их множество как T^W. Запись t^W будет использоваться для обозначения какой–либо одной последовательности транскрипций из T^W.

Применяемые на практике процедуры распознавания речи и обучения акустических моделей, как правило, определяют лучшую последовательность не самих слов, а их акустических моделей, т. е. вместо (1) фактически используется

$$t^{W^*} = \arg\max_{t^W} \frac{P(X|t^W)P(t^W)}{P(X)}. \quad (2)$$

Наиболее вероятная последовательность слов определяется затем путем отнесения каждой модели соответствующему ей слову, т. е.

$$t^{W^*} \to W^*. \quad (3)$$

Когда слова словаря имеют одну единственную транскрипцию, подходы (1) и (2) очевидно эквивалентны.

Используя равенство $P(t^W) = P(t^W|W)P(W)$, выражение (2) можно записать как

$$W^* = \arg\max_{t^W} \frac{P(X|t^W)P(t^W|W)P(W)}{P(X)}. \tag{4}$$

От выражения (2) выражение (4) отличается наличием члена $P(t^W|W)$, который допускает использование вариативности произношения слов. Множество вероятностей $P(T^W|W) = \{P(t^W|W), t^W \in T^W\}$ можно рассматривать как параметры модели вариативности произношения.

Оценка параметров модели вариативности произношения

Для распознавания речи с использованием критерия (4) нужно знать значения параметров трех моделей: акустической, произносительной и модели языка.

Оценка значений параметров по методу максимальной апостериорной вероятности соответствует использованию критерия

$$P(W|X) = \arg\max_{t^W} \frac{P(X|t^W)P(t^W|W)P(W)}{\sum\limits_{t^W \in T^W} P(X|t^W)P(t^W|W)P(W)} \tag{5}$$

Полученные в результате значения параметров можно рассматривать как дискриминантное решение (5) в том смысле, что оно максимизирует вероятность корректных (для обучающих данных) моделей при минимизации суммарной вероятности всех возможных. Параметры модели языка P(W) в (5), как и для (2), можно считать не зависящими от обучающих акустических сигналов, а их оценку выполнять отдельно на текстовом корпусе данных. Однако параметры моделей произношения $P(T^W|W)$, зависят от акустических обучающих данных, поэтому их независимая от параметров акустических моделей оценка некорректна.

Практическое использование критерия (5) по крайней мере до недавнего времени было довольно затруднительным, поскольку предполагало наличие достаточно больших данных, авторам известны пока единичные подобные эксперименты [7]. Для русского языка подобных корпусов данных, по крайней мере опубликованных, пока нет.

Существенно проще выглядит в данной ситуации оценка параметров модели произношения по методу максимального правдоподобия, т. е. с использованием числителя выражения (5).

Предположим, что, обучающий корпус речевых данных X таков, что для всех высказываний известна не только последовательность слов $w_1 w_2 \cdots w_N$ но и их их моделей $t_1^w t_2^w \cdots t_N^w$. В этом случае наиболее правдоподобная оценка параметров $p(t^w|w)$ определится из выражения:

$$p(t^w|w) = \arg\max_{w,t^w} \prod_{w,t^w} p(t^w|w). \tag{6}$$

Эта оценка аналогична соответствующей оценке для вероятностей появления слов в модели языка [12], т. е. это частота появления соответствующей модели:

$$p(t^w|w) = \frac{\#\{t^w\}}{\#\{w\}}, \tag{7}$$

где символ # означает число событий в фигурных скобках, встретившихся в обучающих данных. Таким образом, наиболее правдоподобной оценкой вероятности появления модели слова является ее относительная частота в обучающей выборке.

Поскольку параметры произносительных и акустических моделей очевидно зависят друг от друга, раздельное независимое оценивание их будет некорректно. Предлагается использовать «покоординатную» оптимизацию: сначала получить максимально правдоподобные оценки по одной группе параметров, полагая другие неизменными, потом то же самое для другой группы параметров и т.д.

Например, полагая первоначально все варианты произнесения слов равновероятными, сначала выполнить с использованием существующих акустических моделей распознавание корпуса данных с ограничением на порядок слов, который известен. Вычислить последовательности наиболее вероятных моделей и оценить частоты всех моделей для каждого слова в соответствии с (7). Затем с использованием определенных последовательностей наиболее вероятных моделей обновить значения параметров акустических моделей. Оба этих этапа чередовать до тех пор, пока перестанут изменяться частота появления транскрипций либо вероятность ошибок распознавания.

Алгоритм вычислений значений параметров представлен ниже.

Алгоритм 1 Алгоритм оценки параметров модели произношения

Вход: Речевой корпус данных
текстовая аннотация корпуса данных,
акустические модели аллофонов,
произносительный (фонемный) словарь
Выход: произносительный фонемный словарь

1: **для всех** высказываний корпуса данных:
2: распознаем высказывание
3: определяем наиболее вероятные транскрипции слов
4: вычисляем границы аллофонов и транскрипций
5: **для всех** слов корпуса данных:
6: оцениваем частоты транскрипций
7: **если** частоты транскрипций изменились **то**
8: переоцениваем параметры акустических моделей аллофонов
9: корректируем акустические модели аллофонов
10: корректируем произносительный словарь
11: на шаг 1
12: **иначе**
13: закончить вычисления

Перед началом работы алгоритма каждое слово обучающей части корпуса данных имеет набор транскрипций, которые соответствуют всем практически возможным вариантам его произношения. Оценка эффективности моделей осуществляется по результатам распознавания на независимой тестовой выборке.

Сходимость алгоритма следует из следующих обстоятельств. На шаге распознавания и переоценки частот транскрипций их количество может только уменьшаться за счет отбрасывания редких вариантов. Шаг переоценки параметров моделей при заданных транскрипциях представляет собой модифицированный применительно к марковским моделям

EM-алгоритм (процедура Баума–Уэлча), т. е. правдоподобие данных гарантированно не уменьшается. Хотя теоретические оценки скорости сходимости EM-алгоритма в подобных задачах авторам не известны, практически процесс сходится достаточно быстро, за 4–6 итераций.

Модификация процедур распознавания речи для учета вариативности произношения

Наиболее известный способ реализации модели вариативности произношения при распознавании речи основан на пополнении произносительного словаря новыми вариантами произнесения слов и распознаванием на основе (2)-(3). Этот подход, однако нельзя рассматривать как наилучшее решение.

Перепишем правую часть равенства (1) в виде

$$P(W|X) = \frac{P(W,X)}{P(X)} = \frac{\sum_{t^W \in T^W} P(X, t^W)}{P(X)} = \frac{\sum_{t^W \in T^W} P(X|t^W)P(t^W)}{P(X)}. \tag{8}$$

Из (4) и (8) следует, что если использовать алгоритм вычислений в соответствии с (2), то определить наиболее вероятную последовательность слов W^* можно из условия

$$W^* = \arg\max_W \sum_{t^W \in T^W} P(t^W|X)P(t^W). \tag{9}$$

Решение в соответствии с (9) определяет наиболее вероятную последовательность слов, а не транскрипций, как (2)-(3), что лучше отвечает интуитивному пониманию решения задачи распознавания: как правило нас интересует, какие слова сказаны, а не то, каким образом они были произнесены.

Алгоритм распознавания с использованием критерия (4) отличается от версии для (2)-(3) тем, что нужно учитывать вероятность появления отдельных транскрипций слова, принимать решение о правдоподобии слова по взвешенной сумме правдоподобий его транскрипций.

Реализация вычислений по (9) потребует дополнительные по сравнению с (2)-(3) шаги для выбора лучшей последовательности слов в соответствии с (9). Поскольку теперь для каждого слова w

$$P(w) = \sum_{t^w \in T^w} P(t^w|X), \tag{10}$$

то (если, например, рассматривать произносительный словарь в виде дерева) в каждом листе дерева нужно вычислить вероятность слова в соответствии с (10).

По сравнению с традиционным (2)–(3) подходом также требуется произвести очевидные изменения в структурах данных процедуры поиска, например отвести память для вероятностей отдельных транскрипций слов и связей между словом и листьями дерева, которые соответствуют моделям этого слова.

Практическая реализация этого алгоритма связана с проблемой, которая возникает из-за процедур обрезки (pruning [12]) вершин дерева лексикона при распознавании. Выражение (10) может включать правдоподобия листьев, которые были выброшены из поиска вследствие их малой вероятности. Необходим альтернативный способ оценки значения выражения (10) для таких случаев.

В связи с этим рассмотрим способ оценки правдоподобия слова, упрощенный вариант (10) с заменой взвешенной суммы правдоподобий моделей на выбор взвешенной максимально правдоподобной модели

$$W^* = \arg\max_{W,t^W} P(t^W|X)P(t^W). \tag{11}$$

В этом случае перечисленные выше практические вычислительные проблемы отсутствуют, а сам алгоритм поиска отличается от общепринятого только наличием «штрафующего» члена $P(t^W|X)$.

Эксперименты и обсуждение результатов

Предложенные модели вариативности произношения сравнивались в ходе численного эксперимента на корпусах данных ISABASE–2 [13] и TeCoRus [14].

Необходимым условием экспериментов с вариативностью произношения является наличие частотности у соответствующих слов в корпусе данных: если слово не встречается или встречается в корпусе данных один раз, сложно судить о вариантах его произнесения. Поскольку цифры и числительные достаточно часто повторяются в речи их, в принципе, естественно было бы использовать в таких измерениях. Недостатком является то, что некоторые цифры короткие и их дополнительных вариантов произнесения может и не быть, а также то, что числительные обычно несут смысловую нагрузку, поэтому произносятся аккуратно, что тоже снижает вариативность.

Обучающая выборка состояла из речевых высказываний 200 дикторов ISABASE–2 (около 40 тыс. предложений) и 50 дикторов TeCoRus (3 тыс. предложений). Составленный по обучающей выборке набор числительных включал 26 слов: цифр от 0 до 9 и числительных до сотни. Этот же набор числительных, включая найденные варианты использовался во всех описанных ниже тестах.

Материал первого теста состоял из данных TeCoRus: 776 слитных цифровых последовательностей, длиной от 2 до 16 цифр, всего 3147 цифр, фактический словарь соответственно был ограничен словоформами цифр. Чтобы оценивать влияние только акустических характеристик на распознавание, модель языка была отключена.

Результаты численного эксперимента, выраженные в терминах пословной ошибки распознавания WER приведены в табл. 1. Здесь и далее в приведенных таблицах колонка «Базовый» содержит результаты для случая использования только базовых фонемных транскрипций, т. е. без вариативности произношения: каждое слово имеет ровно одну, базовую, фонемную транскрипцию . Колонка «Обычный» соответствует методу учета вариативности с использованием (1)–(3), колонка «Опт» – методу учета вариативности на основе (9) и колонка «СубОпт» — методу на основе (11). Значение в строке «Вариат.» характеризует среднюю вариативность произносительного словаря, т. е. среднее число вариантов произносительных транскрипций на слово, которое оценивалась как $\sum_{i=1}^{N} t_i/N$ где N – количество слов в словаре, а t_i – число произносительных транскрипций которые имело i-е слово. В качестве словаря, для которого вычислялась средняя вариативность, рассматривался словарь цифр и числительных.

Результаты приведенные в табл. 1 можно интерпретировать как свидетельство отсутствия вариативности произнесения цифр в данном корпусе. Действительно, дикторы TeCoRus проживали в Москве, имели высшее образование, принадлежали в основном к двум профессиональным группам (в том числе лингвисты), которые аккуратно читали предложенный цифровой материал.

Таблица 1. Показатель пословной ошибки распознавания (WER) для моделей вариативности произношения на данных корпуса TeCoRus

Метод	Базовый	Обычный	Опт	СубОпт
WER	1,62	5,78	2,00	3,17
Вариативный	1,0	1,9	1,9	1,9

Второй эксперимент был проведен на предположительно более вариативном тестовом материале. Обучающий материал был тем же, что и в первом тесте. Тестовый материал включал весь материал первого теста, а также высказывания дикторов TeCoRus, которые содержали числовую информацию (даты, время, номера и т.п.): всего 867 последовательностей цифр или высказываний от 11 дикторов корпуса TeCoRus. Словарь тестовой части включал 129 словоформ, включая 26 цифр и числительных с потенциальной вариативностью. Записи тестовой части, относящиеся к высказываниям с числовой информацией включали потенциально устную речь: дикторы отвечали на вопрос, например, о номерах школ, датах рождения, почтовых индексах и другой персональной числовой информации. Записи в данном случае также включали существенное число различных речевых нарушений. Их было сложно удалить без искажения сигнала. Они служили одним из источников ошибок распознавания.

Таблица 2 содержит результаты измерений уровня пословной ошибки WER в этом случае.

Таблица 2. Показатель WER при использовании различных способов учета вариативности произношения для числовых данных TeCoRus

Метод	Базовый	Обычный	Опт	Субопт
WER	7,78	7,57	7,38	7,44
Вариативный	1,0	1,3	1,3	1,3

Результаты, приведенные в (2), можно рассматривать как более ожидаемые: оптимальным для минимизации показателя WER оказалось использование метода частотного взвешивания произносительных вариантов (9). Метод простого добавления транскрипций (1)–(3) оказался менее эффективным по сравнению как с оптимальным, так и субоптимальным (11), которые учитывают частотность транскрипций, но все же предпочтительнее, чем использование только канонических моделей.

В то же время изменения показателя WER в результате использования моделей произношения представляются незначительными.

Третий эксперимент был проведен на материале естественной разговорной речи. Корпус данных был специально собран и аннотирован для этого эксперимента. Он состоял из фрагментов интервью, которые были взяты с сети Интернет, с сайта радиостанции «Эхо Москвы» [15]. Аудио фрагменты были конвертированы из формата MP3 в формат WAV. Речь была естественная, разговорная в нормальном или быстром темпе. Интервью были предварительно автоматически сегментированы по репликам, найдены и выделены в отдельные файлы фрагменты, которые содержали цифры и числительные.

Полученная таким образом тестовая выборка состояла из 200 коротких реплик с общим словарем в 91 слово, включающим в себя и описанный выше словарь числительных с вариантами произнесения.

Таблица 3 содержит значения показателя WER для этого теста.

Таблица 3. Значения WER для различных моделей вариативности произношения для материала с разговорной русской речью

Метод	Базовый	Обычный	Опт	СубОпт
WER	69,3	57,44	59,7	60,0
Вариативный	1,0	1,3	1,3	1,3

Существенно более высокий уровень абсолютной ошибки распознавания вызван характером речи, использованием кодека, отключением модели языка и несоответствием акустических моделей обучающих и тестовых данных. Кроме того, процедура распознавания и сегментации, которая выделяла словосочетания с числительными из слитной речи не всегда корректно определяла границы слов при слитном произношении, например, в словосочетаниях.

В отличие от предыдущих тестов, где уменьшение относительного уровня ошибки не превышало 5%, в данном случае наблюдается уменьшение относительного уровня ошибок от 13,4% до более чем на 17,1%.

Такие образом для ситуации со словами и словосочетаниями в естественной речи использование моделей вариативности произношения приводит к существенному снижению уровня ошибок распознавания.

Помимо фактически наблюдаемой вариативности произношения возможной причиной снижения уровня ошибок может быть также несоответствие акустических характеристик обучающего и тестового материала, например, из–за использования кодека. За счет этого параметры моделей могли измениться настолько, что это отразилось на наблюдаемых фонемных транскрипциях. Однако это обстоятельство вряд ли играет в данном случае важную роль, поскольку во втором тесте такого эффекта не наблюдалось.

Оптимальный метод формально не оказался лучшим в данном тесте. Этому эффекту можно дать объяснение. Взвешивание решений по транскрипциям уравнивает частоты появления слов. В данном тесте числа имеют в тестовом материале гораздо бóльшую частоту появления, чем другие слова. Использование простого метода пополнения транскрипциями соответствует бóльшей частоте появления для чисел, что соответствует фактическим данным. Таким образом, существенного отклонения от теоретически ожидаемого поведения методов в результатах теста нет.

Заключение

Выполнено исследование методов повышения точности автоматического распознавания русской речи за счет использования явных моделей вариативности произношения. Определена вероятностная модель вариативности произношения, даны способы вычисления ее параметров и способы включения модели вариативности в процедуры распознавания речи. Выполнены численные эксперименты и установлено, что:

— при распознавании т.п. подготовленной читаемой русской речи использование моделей вариативности не является эффективным способом повышения точности распознавания, даже может ухудшать показатели распознавания;

— при распознавании русской спонтанной разговорной речи, в том числе при несоответствии характеристик обучающего и тестового речевого материала, модели вариативности являются эффективным способом снижения уровня ошибок;

— для успешного применения оптимальных моделей вариативности, которые учитывают частоты появления вариантов транскрипций, нужно иметь адекватный тестовому обучающий материал, в противном случае выигрыш от использования таких моделей, по сравнению с традиционными, отсутствует;

Литература

[1] *Fosler-Lussier E.* Dynamic pronunciation models for automatic speech recognition. Ph.D. Thesis. Berkley, CA: University of California, 1999.

[2] *Wester M.* Pronunciation modeling for ASR — knowledge-based and data-derived methods // *Computer Speech Language*, 2003. Vol. 17. P. 69–85.

[3] *Saraclar M., Khudanpur S.* Pronunciation change in conversational speech and its implications for automatic speech recognition // *Computer Speech Language*, 2004. Vol. 18. No. 4. P. 375–395.

[4] *Word Error Rate* `http://www.echo.msk.ru`.

[5] *Saraclar M., Nock H., Khudanpur S.* Pronunciation modeling by sharing Gaussian densities across phonetic models // *Computer Speech Language*, 2000. Vol. 14. No. 4. P. 137–160.

[6] *Zheng J., Franco H., Stolcke A.* Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition // *Speech Communication*, 2003. Vol. 41. P. 273–285.

[7] *Lehr M., Gorman K., Shafran I.* Discriminative pronunciation modeling for dialectal speech recognition // *Proc. Interspeech*, 2014 (in press).

[8] *Hitchinson B., Droppo J.* Learning non-parametric models of pronunciation in automatic speech recognition // *Conference (International) on Acoustics, Speech, and Signal Processing, ICASSP, Proceedings*, 2011. P. 4904–4907.

[9] *Hain T.* Implicit modelling of pronunciation variation in automatic speech recognition // *Speech Communication*, 2005. Vol. 46. P. 171–188.

[10] *Bahl R., Jelinek F., Mercer R. L.* A maximum likilihood approach to continuous speech recognition // *IEEE Trans. Pattern Anal. Machine Intell.*, 1983. Vol. 5. P. 179–190.

[11] *Jelinek F.* Statistical methods for speech recognition. Cambridge, MA: The MIT Press, 1997.

[12] Corpus-based methods in language and speech processing / Ed. by S. Young, G. Bloothooft. Dordrecht: Kluwer Academic Publishers, 1997.

[13] *Богданов Д. С., Кривнова О. Ф., Подрабинович А. Я., Арлазаров В. Л.* Creation of Russian speech databases: Design, processing, development tools // *Conference (International) on Speech and Computers, SPECOM, Proceedings.* Москва, 2004.

[14] *Чучупал В.Я., Маковкин К.А, Чичагов А.В., Кузнецов В.Б., Огарышев В.Ф.* Речевой корпус данных TeCoRus. Свидетельство об официальной регистрации базы данных № 2005620205, 2005.

[15] *Сайт радиостанции Эхо Москвы.* `http://www.echo.msk.ru`.

References

[1] *Fosler-Lussier, E.* 1999. Dynamic pronunciation models for automatic speech recognition. Ph.D. Thesis. Berkley, CA: University of California.

[2] *Wester M.* 2003. Pronunciation modeling for ASR — knowledge-based and data-derived methods. *Computer Speech Language* 17:69–85.

[3] *Saraclar M., Khudanpur S.* 2004. Pronunciation change in conversational speech and its implications for automatic speech recognition. *Computer Speech Language* 18(4):375–395.

[4] *Word Error Rate* `http://www.echo.msk.ru`.

[5] *Saraclar M., Nock H., Khudanpur S.* 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. *Computer Speech Language* 14(4):137–160.

[6] *Zheng J., Franco H., Stolcke A.* 2003. Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition. *Speech Communication* 41:273–285.

[7] *Lehr M., Gorman K., Shafran I.* 2014 (in press). Discriminative pronunciation modeling for dialectal speech recognition. *Proc. Interspeech.*

[8] *Hitchinson B., Droppo J.* 2011. Learning non-parametric models of pronunciation in automatic speech recognition. *Conference (International) on Acoustics, Speech, and Signal Processing, ICASSP, Proceedings.* 4904–4907.

[9] *Hain T.* 2005. Implicit modelling of pronunciation variation in automatic speech recognition. *Speech Communication* 46:171–188.

[10] *Bahl R., Jelinek F., Mercer R. L.* 1983. A maximum likilihood approach to continuous speech recognition. *IEEE Trans. Pattern Anal. Machine Intell.* 5:179–190.

[11] *Jelinek F.* 1997. Statistical methods for speech recognition. Cambridge, MA: The MIT Press.

[12] *Young S., and G. Bloothooft, eds.* 1997. Corpus-based methods in language and speech processing. Dordrecht: Kluwer Academic Publishers. 235 p.

[13] *Bogdanov D. S., Krivonva O. F., Podrabinovitch A. J., Arlazarov V. L.* 2004. Creation of Russian speech databases: Design, processing, development tools. *Conference (International) on Speech and Computers, SPECOM, Proceedings.* Moscow. (in Russ.)

[14] *Chuchupal V. J., Makovkin K. A. Chichagov A. V., Kouznetsov V. B., Ogaryshev V. F.* 2005. Speech data corpus TeCoRus. Federal Institute of Industrial Property, RosPatent Database register No. 2005620205. (in Russ.)

[15] *"Echo of Moscow" News/Media WebCite.* `http://www.echo.msk.ru`.