Сравнение эффективности моделей вариативности произношения для систем распознавания речи[1]

Чучупал В.Я., *в.н.с. ВЦ РАН, кандидат физико-математических наук*

Коренчиков А.А., *студент 5 курса МГУ*

Работа посвящена исследованию способов моделирования вариативности произношения в системах распознавания речи. Определена вероятностная модель произношения слова, приведены алгоритмы оценки ее параметров. Приведено сравнение, с точки зрения уровня ошибок и сложности реализации, нескольких вариантов реализации модели вариативности произношения в системе распознавания речи.

• автоматическое распознавание речи • вариативность речи • моделирование произношения • скрытые марковские модели.

The paper addresses the problem of pronunciation modeling for automatic speech recognition. A statistical pronunciation model based on the explicit approach is reviewed along with the ways for estimation of its parameters and its implementation in the speech recognition engines. The experimental results are described that show the usefulness of the proposed approach in terms of WER gain.

[1] Работа выполнена при финансовой поддержке РФФИ, проект 11-01-00900а.

Введение

Акустический образ слова в системах распознавания речи обычно моделируется путём задания его фонемной (или произносительной) транскрипции как последовательности составляющих это слово фонем. Большинство слов в произносительном словаре систем распознавания речи представлено единственной транскрипцией, которая соответствует каноническому (нормативному или базовому) произнесению.

Вариативность в произношении слов является одной из основных причин появления ошибок при распознавании речи. Она может возникать вследствие различных обстоятельств: индивидуальных особенностей, темпа речи, эмоционального состояния говорящего и т.п. При автоматическом распознавании к перечисленным выше причинам может добавиться вариативность, вызванная неадекватностью акустических моделей наблюдаемым данным из-за различия между обучающим и фактическим материалом.

По приведенным выше причинам конкретное произнесение слова может существенно отличаться от нормативного, что часто приводит к ошибкам в его распознавании.

Под моделированием вариативности произношения в речевой технологии подразумевают набор моделей и методов для определения множества наиболее вероятных в той или иной ситуации акустических образов слов и словосочетаний.

В литературе встречаются два основных подхода к моделированию вариативности произношения [2, 3]. Явное моделирование (explicit modeling) заключается в моделировании вариативности произношения путем модификаций фонемных транскрипций слов [2]. При неявном моделировании (implicit modeling) [4] вариативность произношения обеспечивается путем изменений структуры моделей звуков канонической транскрипции.

Оба подхода никоим образом не отменяют использования канонических транскрипций и направлены на определение дополнительных вариантов произнесения слов и словосочетаний.

В данной работе мы следовали явному подходу к моделированию вариативности произношения, то есть, предполагали, что все наблюдаемые изменения в произношении можно адекватно описать соответствующими изменениями фонемных транскрипций.

Практическая реализация такого подхода для моделирования вариативности произношения слов в системе распознавания речи обычно связана с решением следующих задач:

- выбор модели вариативности и ее параметров;
- определение наиболее вероятных вариантов произнесения слов;
- определение алгоритма использования вариантов произнесения при распознавании.

Соответственно, в работе рассмотрена вероятностная модель вариативности произношения, алгоритмы оценки ее параметров и реализации в процедурах поиска. Приведено сравнение эффективности использования нескольких вариантов моделей вариативности при распознавании цифр и числительных.

Модель вариативности произношения

Цель использования модели вариативности произношения в системе распознавания речи — уменьшение числа ошибок в распознавании. При использовании явного подхода это предполагается достичь за счет использования кроме базовых транскрипций, их вариантов, которые более соответствуют фактическим произнесениям.

Для того, чтобы показать, как можно эффективно использовать вариативность произношения, напомним формулировку вероятностного подхода к распознаванию речи [5].

Пусть $X = \{x_t | t = 1,\ldots,T\}$ — последовательность параметров наблюдаемого речевого сигнала, а $W = \{w_i | i = 1,\ldots,N\}$ — последовательность слов словаря. Результат распознавания образа X, наиболее вероятная последовательность слов W^*, определится путем оптимизации выражения [1]:

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

Первый сомножитель — $P(X|W)$ в числителе (1) соответствует правдоподобию данных при заданной последовательности слов. Полученная величина правдоподобия затем умножается на второй сомножитель, значение $P(W)$, определяемое с помощью модели языка. Отметим, что моделей произношения слов, транскрипций, в (1) в явном виде нет.

Различие между словами и транскрипциями заключаются в том, что слова относятся к смыслу высказывания и записываются в орфографической форме, а их произносительные транскрипции определяют акустические параметры и образы слов. Это различие можно учесть, дополнив критерий (1) моделью вариативности произношения.

Пусть акустической моделью некоторого слова W служит его произносительная транскрипция t^w. Множество всех транскрипций слова w обозначим T^w. Моделью последовательности слов W может быть любая последовательность их транскрипций. Обозначим это множество как T^W. Запись t^W будет использоваться для обозначения какой-либо одной последовательности транскрипций из множества T^W.

Отметим, что применяемые на практике процедуры распознавания речи фактически определяют лучшую последовательность не самих слов, а их произносительных транскрипций [6], т.е. вместо (1) при распознавании оптимизируется:

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

Наиболее вероятная последовательность слов определяется затем путем отнесения каждой произносительной модели в последовательности t^{W*} соответствующему ей слову:

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

Поскольку на практике слова из словаря, как правило, имеют одну единственную транскрипцию, отображение (3) однозначно и критерии (1) и (2) эквивалентны. При фактическом наличии вариативности произношения эти критерии, очевидно, уже не будут эквивалентны.

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

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

Запись в форме (4) позволяет явно отделить, помимо акустической модели и модели языка, как оценки вероятности наблюдения набора слов $P(W)$, также модель вариативности произнесения, как вероятности появления заданной последовательности транскрипций $P(t^W|W)$ для данной последовательности слов словаря. Множество вероятностей $\{P(t^W|W)\}$ естественно при этом рассматривать как параметры такой модели вариативности.

Оценка параметров модели вариативности произношения

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

Оптимальная оценка значений параметров моделей по методу максимальной апостериорной вероятности соответствует использованию критерия:

$$P(W \mid X) = \arg\max_{t^W} \frac{P(X \mid t^W)P(t^W \mid W)P(W)}{\sum_{t^W} P(X \mid t^W)P(t^W \mid W)P(W)}, \quad (5)$$

где используется, что $P(X) = \sum_{t^W} P(X \mid t^W)P(t^W \mid W)P(W)$.

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 $P(W)$ 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 X таков, что известна последовательность слов $w_1 w_2 \dots w_N$ и их моделей — транскрипций: $t^w_1\ t^w_2 \dots t^w_N$, которая соответствует речевым высказываниям в X. В предположении, что последовательность транскрипций не изменяется в процессе обучения акустических моделей, наиболее правдоподобная оценка параметров модели произношения $p(t^w|w)$ определится из:

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

Решение (6) получается совершенно аналогично соответствующим оценкам для вероятностей появления слов в модели языка [7], т.е это соответствующие частоты встречаемости:

$$p(t^w \mid w) = \frac{\#\{t^w\}}{\#\{w\}}, \quad (7)$$

где символ # означает число событий, встретившихся в обучающих данных.

Таким образом, наиболее правдоподобная оценка вероятности появления транскрипции слова равно её относительной частоте в обучающей выборке.

Поскольку параметры произносительных и акустических моделей очевидно зависят друг от друга, раздельное независимое оценивание их, по (6), в отличие от параметров модели языка, некорректно.

В этом случае предлагаем использовать алгоритм попеременной оптимизации: сначала получить оптимальные оценки по одной группе параметров полагая другие неизменными, а затем сделать то же самое для другой группы параметров. Более конкретно, предполагая вначале все варианты произнесений слов равновероятными, т.е. $p(t^w_i \mid w) = p(t^w_j \mid w), i \neq j$, выполним, с использованием существующих акустических моделей, распознавание фраз из корпуса данных. Вычислим последовательности наиболее вероятных фактических транскрипций и получим оценку их частот в соответствии с (7). Далее вычислим, для только что определенной последовательности транскрипций, новые значения параметров акустических моделей и выполним заново

распознавание фраз из корпуса. Оба этапа (оценки частот транскрипций и параметров акустических моделей) чередуем до тех пор, пока перестанут меняться либо частоты появления транскрипций, либо вероятность ошибок распознавания.

Принципиальная блок-схема соответствующего алгоритма приведена на следующем рис. 1.

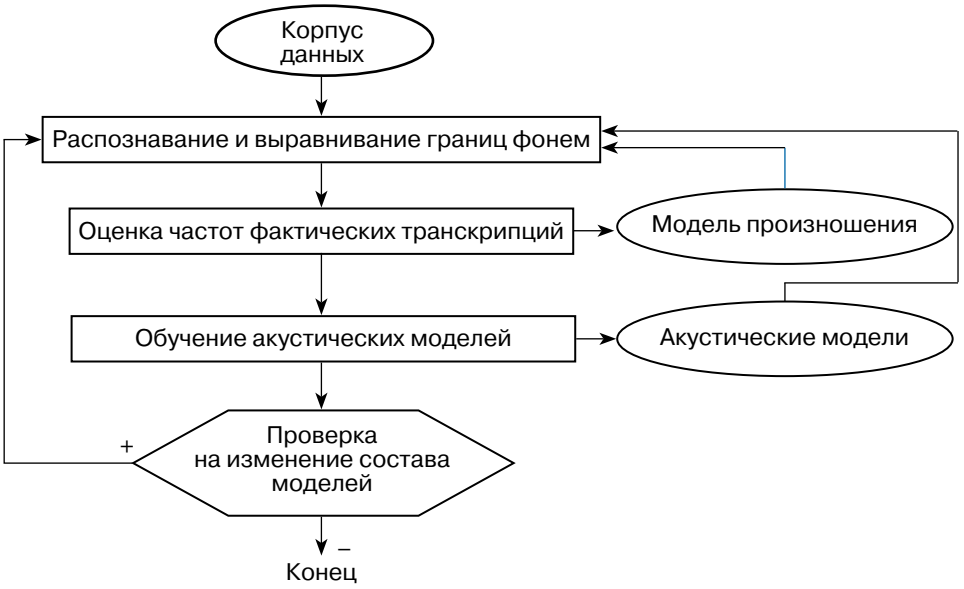

Рис 1. Алгоритм оценки параметров модели произношения

Отметим, что перед началом работы алгоритма каждое слово из обучающей части корпуса данных имеет набор потенциально допустимых транскрипций, которые отвечают возможным вариантам его произнесения.

Оценка эффективности полученных таким образом моделей вариативности произношения осуществляется по результатам распознавания на тестовой выборке.

Модификация процедур распознавания речи для учёта вариативности произношения

Наиболее известный и простой способ реализации вариативности произношения при использовании (2) основан на пополнении произносительного словаря новыми вариантами произнесения, наравне с каноническими транскрипциями и нахождением решения в соответствии с выражениями (2)–(3).

Как уже отмечалось, недостатком такого решения является то, что в этом случае фактически определяется наиболее вероятная последовательность транскрипций слов, но не самих слов.

Формальное условие для определения наиболее вероятной последовательности слов W^* можно получить, если записать правую часть равенства (1) в виде:

$$P(W \mid X) = \frac{P(W,X)}{P(X)} = \frac{\sum_{T \in T^W} P(X,T)}{P(X)} = \frac{\sum_{T \in T^W} P(X \mid T)P(T)}{P(X)}. \qquad (8)$$

Из (4) и (8) следует, что наиболее вероятная последовательность слов может быть получена как:

$$W^* = \arg\max_W \sum_{T^W} P(T^W \mid X)P(T^W). \qquad (9)$$

Решение в соответствии с (9) определяет наиболее вероятную последовательность слов, а не транскрипций, что лучше отвечает интуитивному пониманию решения задачи распознавания: важнее, 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 принимать решение о правдоподобии слова теперь нужно по взвешенной сумме правдоподобий его транскрипций.

Реализация вычислений по (9) достаточно очевидна, но требует дополнительные, по сравнению с обычным (2)–(3) алгоритмом, шаги.

Несмотря на теоретическую оптимальность практическая реализация алгоритма на основе (9) связана с проблемой, которая возникает из-за процедур обрезки (pruning, [7]) вершин дерева лексикона при распознавании. Для того, чтобы поиск мог быть выполнен в разумные по времени сроки все вершины дерева, которые имеют невысокое правдоподобие, выбрасываются из дальнейшего поиска, обрезаются. Таким образом, для практического использования нужно предложить вариант вычислений (9) в тех случаях, когда листья лексикона (предположим, без ограничения общности, что словарь представлен в виде префиксного дерева)обрезаны вследствие их малой вероятности. В этом случае их правдоподобие неизвестно и нужно модифицировать алгоритм (9), что в любом случае приводит к потере его оптимальности.

В экспериментах, описанных в следующем разделе, при использовании алгоритма вычислений в соответствии с (9), в тех случаях, когда листья оказывались обрезаны, правдоподобие соответствующих транскрипций аппроксимировалось значением текущего порога обрезки.

Для преодоления описанных недостатков рассмотрим следующий способ оценки правдоподобия слова, неоптимальный вариант (9) с заменой взвешенной суммы правдоподобий моделей на выбор одной максимально правдоподобной модели с весом в виде:

$$W^* = \arg\max_{W,T^W} P(T^W \mid X)P(T^W) \qquad (10)$$

Алгоритм вычислений в соответствии с (10) избавлен от вышеупомянутой проблемы и фактически отличается от (2)–(3) только наличием «штрафующего» множителя $P(T^W \mid X)$, т.е. обладает такой же вычислительной сложностью.

Численные эксперименты

Эффективность трех рассмотренных выше алгоритмов: на основе (2)–(3), (9) и (10) оценивалась в ходе численного эксперимента, который выполнялся на корпусах данных ISABASE-2 [8] и TeCoRus [9] на материале, который в основном состоял из цифр и чисел. Обучащая выборка включала речевые высказывания 200 дикторов ISABASE-2 (40000 предложений) и 50 дикторов TeCoRus (3000 предложений) , тестовая — 776 предложений (3147 цифр) от 11 дикторов TeCoRus.

Таким образом, в экспериментах использовалась, в основном, вариативность цифр и числительных. Словарь включал 130 слов, число появлений цифр в корпусе на три по-

рядка больше, чем других слов. Записи также включали небольшое число «неслов»: оговорок, запинок и т.п., которые порождали дополнительные ошибки.

Отметим, что использовать в экспериментах только данные TeCoRus, который содержал значительное количество цифр и числительных было не совсем репрезентативно, так как дикторы TeCoRus принадлежали в основном к одной профессиональной и локализованной по месту жительства группы, говорили достаточно медленно и аккуратно, то есть были основания предполагать, что в данном случае заметной вариативности произнесения (цифр и числительных) может не оказаться.

Результаты численного эксперимента по сравнительной эффективности распознавания для всех трех алгоритмов (только на TeCoRus и только для цифр т.е. словарь из 10 слов) приведены в Табл. 1. В качестве меры уровня ошибок использовался широко известный показатель пословной ошибки распознавания WER (word error rate). Здесь колонка «Обычный» соответствует методу (2)–(3), «Оптимальный» — методу (9) и «Субопт.» — методу (10). Вариативность произношения определялась как среднее число фактических транскрипций, которые приходились на одно слово словаря.

Табл.1

Значения показателя пословной ошибки распознавания при использовании различных вариантов учета вариативности произнесения (только на данных TeCoRus)

Метод	Обычный		Оптимальный	Субоптим.
Ошибка WER	1.62	5.78	2.00	3.17
Вариативность	1.0	1.9	1.9	1.9

Эти результаты можно интерпретировать как свидетельство фактического отсутствия вариативности произнесения цифр в данном корпусе, что, как было указано выше, вполне объяснимо. Это согласуется и с поведением алгоритма обучения: если оценивать варианты только на TeCoRus, то с увеличением итераций в алгоритме Рис. 1 среднее количество вариантов на слово приближается к 1.

На следующей Табл. 2 показаны результаты измерений показателя WER (на том же тестовом материале) для вариантов произношения и моделей, которые оценивались на основной выборке — корпусах данных ISABASE-2 и TeCoRus.

Табл. 2

Значения показателя пословной ошибки распознавания при использовании различных способов учета вариативности произношения

Метод	Обычный		Оптимальный	Субоптим.
Ошибка WER	7.78	7.57	7.38	7.44
Вариативность	1.0	1.3	1.3	1.3

Результаты, приведенные в Табл.2, можно считать соответствующими теоретическим, поскольку оптимальным для минимизации показателя WER оказалось использование метода частотного взвешивания вариантов произнесений (9). Метод простого добавления транскрипций (2)–(3) оказался

менее эффективным, по сравнению с как с оптимальным, так и субоптимальным (10), которые учитывают частотность транскрипций, но все же предпочтительнее, чем использование только канонических моделей.

Изменения показателя WER в результате использования моделей произношения были невелики, предполагаем, что это существенно зависит от словаря (цифры, например, нельзя назвать вариативными, они как именованные сущности, обычно произносятся достаточно разборчиво) и от условий, в которых осуществляется речевая коммуникация. В данном случае оба корпуса данных были записаны в Москве хорошо образованными дикторами, материал — читаемый, то есть условия возникновения вариативности в значительной мере отсутствовали.

Заключение

Рассмотрены вопросы практической реализации методов моделирования произношения в системах распознавания речи. В частности предложены алгоритмы для оценки параметров моделей произношения и алгоритмы поиска с использованием этих моделей, которые позволяют вычислить результаты распознавания с использованием критерия лучшей последовательности слов по сравнению с критерием определения лучшей последовательности состояний, который обычно используется при реализации поиска на основе алгоритма Витерби.

Список литературы

1. *Jelinek F*. Statistical Methodsfor Speech Recognition // The MIT Press, Cambridge, Massachusetts, 1997.

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

3. Fosler-Lussier E. Dynamic pronunciation models for automatic speech recognition // Ph.D. thesis, University of California, Berkley, CA, 1999.

4. *Saraclar M., Khudanpur S*. Pronunciation change in conversational speech and its implications for automatic speech recognition // Computer Speech and Language, Vol.18, Issue 4, 375–395, 2004.

5. *Bahl L.R., Jelinek F., Mercer R.L*. «A maximum likelihood approach to continuous speech recognition», IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-5, pp.179–190, 1983.

6. *Chow Y.-L., Richard Schwartz R*. The N-Best Algorithm: Efficient Procedure for Finding Top N.

7. Sentence Hypotheses. // Proc. Int. Conf on Acoustic, Speech and Signal Processing, ICASSP, 1990, Pp. 199–202.

8. *Young S., Bloothooft G.,* editors. Corpus-based methods in language and speech processing // Text, Speech and Language Technology, Vol. 2, Kluwer Academic Publishers, 1997.

9. *Богданов Д.С., Кривнова О.Ф.,* 11. *Кривнова О.Ф., Богданов Д.С., Подрабинович А.Я., Арлазаров В.Л*. Creation of Russian Speech Databases: Design, Processing, Development Tools // International Conference SPECOM'2004. Proceedings. СПб. 2004.

10. *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 г.