УДК 004.934

Г. А. Черных, М. Л. Кореневский, К. Е. Левин,
И. А. Пономарева, Н. А. Томашенко

КРОССВАЛИДАЦИОННЫЙ КОНТРОЛЬ СОСТОЯНИЙ ПРИ ОБУЧЕНИИ АКУСТИЧЕСКИХ МОДЕЛЕЙ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Предложен метод, позволяющий при обучении скрытых марковских моделей, входящих в состав систем автоматического распознавания речи, оптимизировать число компонент в гауссовых смесях состояний. Применение метода повышает качество и скорость распознавания речи системой.

Ключевые слова: *автоматическое распознавание речи, скрытые марковские модели, кроссвалидационный контроль, гауссова смесь.*

Введение. Скрытые марковские модели (hidden Markov models, HMM) являются важной составляющей многих систем автоматического распознавания речи [1]. При обучении HMM-моделей часто возникает проблема недостаточного количества обучающих данных. В подобных случаях необходимо соблюдать баланс между сложностью моделей и их обобщающей способностью. Чрезмерно сложные модели, содержащие множество гауссиан в гауссовых смесях состояний, склонны к переобучению и, как следствие, теряют свою эффективность на данных, не входящих в обучающую выборку.

Предлагаемый в настоящей работе подход позволяет оптимизировать сложность моделей в соответствии с речевыми данными, в результате чего снижается вычислительная сложность алгоритма и повышается качество распознавания речи. Этот подход был успешно применен при обучении акустических моделей системы автоматического распознавания речи ООО „ЦРТ“ для русского языка, основой которой является тандемная структура [2], где сформированные нейронными сетями и подвергнутые дополнительной обработке [3] акустические признаки речи используются для обучения HMM-моделей.

Кроссвалидационный контроль используется во многих задачах машинного обучения с целью исключения эффекта переобучения, например, при обучении искусственных нейронных сетей в ходе подстройки параметров. Однако обучение HMM-модели сопровождается усложнением самого субъекта (модели) в результате последовательного расщепления гауссовых смесей. Кроссвалидационный контроль можно применять и при обучении HMM-моделей, но о его использовании в таких задачах практически не упоминается в литературе; в качестве критерия останова процесса обучения традиционно предлагается проводить тестирование качества распознавания на небольшой выборке [4]. Представленный в настоящей работе метод позволяет эффективно решать задачу построения гауссовых смесей и своевременно останавливать обучение в целом.

Метод основан на использовании специальных кроссвалидационных критериев, регулирующих количество компонент в гауссовых смесях состояний при итеративном обучении моделей. Критерии вводятся для каждого состояния в отдельности и строятся по отдельной части речевой базы, не принимающей участия в обучении моделей. Обучение HMM с увеличением количества гауссиан в состояниях заключается в чередовании эпох* „спокойного“ обучения, когда происходит только подстройка параметров без усложнения моделей, и одномоментных расщеплений гауссиан состояний [3]. Решение о целесообразности расщепления

* Эпоха (*англ.* epoch) — однократный проход по всему обучающему множеству.

каждого из состояний принимается на основе динамики кроссвалидационных критериев в конце эпохи обучения (непосредственно перед моментом очередного расщепления). Таким образом, количество гауссиан увеличивается не для всех, а только для отобранных состояний. Кроме того, в ходе обучения возможно и ухудшение динамики критерия в некоторых состояниях, например, вследствие эффекта переобучения, вызванного недостатком данных. В этом случае происходит возврат состояний в их предыдущие реализации, для которых соответствующие кроссвалидационные критерии показывали бóльшую эффективность обучения.

Стандартная процедура обучения HMM-модели состоит из следующих основных этапов (см., например, [4]):

1) обучение монофонных одногауссовых моделей;

2) получение несвязанных одногауссовых трифонных моделей посредством клонирования монофонных моделей;

3) дообучение несвязанных одногауссовых трифонных моделей;

4) связывание состояний одногауссовых трифонных моделей с использованием дерева решений;

5) обучение трифонных моделей со связанными состояниями с последовательными расщеплениями состояний и получение многогауссовых трифонных моделей [5].

Предлагаемый в настоящей работе метод может применяться на пятом этапе обучения, во время которого происходит последовательное наращивание количества гауссиан в связанных состояниях трифонных моделей.

Прежде чем перейти к описанию алгоритма, отметим две особенности предлагаемого подхода. Во-первых, сбор статистики для обновления акустических моделей производится с помощью разновидности алгоритма Баума—Уэлша, так называемого алгоритма „точного совпадения“ (exact-match), используемого при дискриминативном обучении [6]. В отличие от канонического алгоритма Баума—Уэлша, требующего только информацию о последовательностях фонем, версии „точного совпадения“ необходима пофонемная разметка обучающих данных [7]. При этом алгоритм „прямого-обратного“ хода (forward-backward) [1], который является основой метода сбора статистики для обновления HMM-моделей, применяется не к каждой фразе целиком, а к интервалам, соответствующим положению отдельных фонем этой фразы в разметке (при дискриминативном обучении эти интервалы соответствуют ребрам фонемной сети гипотез). Знание относительных вероятностей посещения (occupancy probability) [1] на интервалах, их временных границ и последовательности фонем позволяет получить абсолютные значения вероятностей, используемых для обновления моделей. Подобный подход значительно ускоряет процедуру обучения и в общем случае улучшает качество моделей, так как в процессе разметки из обучающего множества могут быть исключены заведомо ошибочные данные. Во-вторых, поскольку алгоритм призван оптимизировать количество гауссиан, добавляется только по одной гауссиане в состояние при его расщеплении.

Кроссвалидационный контроль и его критерии. При обучении моделей речевые данные разделяются на две части: первая (обучающая) применяется непосредственно для обучения, вторая (кроссвалидационная) используется только для вычисления критериев. Процедура вычисления кроссвалидационного критерия для состояния состоит в следующем. К кроссвалидационным данным, аналогично тому как это делается на обучающих данных с учетом специфики подхода „точного совпадения“, применяется алгоритм „прямого-обратного“ хода. Для каждого состояния s HMM-модели, присутствующего в последовательности фонем обрабатываемой фразы, вычисляются вероятности его наблюдения в каждый из моментов времени, соответствующих фонеме. Набор вероятностей наблюдения образует кусочную функцию дискретного времени $L_i^s(t)$, где i — номер фразы. Из этих значений вероятности выбираются все локальные максимумы $\hat{t}_1^s,\ldots,\hat{t}_{M(s,i)}^s$ и помещаются в „аккумулятор“

состояния. После обработки всех фраз, входящих в кроссвалидационную выборку, вычисляется среднее по „аккумулятору“ состояния, что и является искомым критерием:

$$C(s) = \frac{\sum_{i=1}^{N} \sum_{k=1}^{M(s,i)} L_i^s\left(\hat{t}_k^s\right)}{\sum_{i=1}^{N} M(s,i)}, \quad (1)$$

где N — число фраз в кроссвалидационной выборке.

Вместе с индивидуальными кроссвалидационными критериями состояний, позволяющими оптимизировать количество гауссиан в состояниях, вводится общий кроссвалидационный критерий, который используется для фиксации времени останова обучения. Критерий вычисляется усреднением значений правдоподобия, определяемых с помощью алгоритма „прямого-обратного“ хода по полным фразам кроссвалидационной выборки. Пусть T_i — число кадров в i-й фразе кроссвалидационной выборки и LLH_i — суммарная вероятность, тогда критерий:

$$C = \frac{1}{N} \sum_i \frac{\mathrm{LLH}_i}{T_i}. \quad (2)$$

Поскольку обучение происходит по критерию максимального правдоподобия, то значения аналогичных критериев, вычисляемые на основе обучающей выборки, в процессе обучения должны возрастать. Однако динамика критериев, рассчитанных по кроссвалидационной выборке, в ходе обучения и наращивания количества гауссиан может быть довольно сложной и необязательно монотонной. Описываемый далее алгоритм расщепления гауссовых смесей эффективно обеспечивает рост критериев в целом.

Стратегия расщепления гауссовых смесей. Предлагаемый алгоритм расщепления характеризуется следующими основными особенностями:

— слежение за каждым состоянием в отдельности;

— использование резервных копий состояний для возврата к более удачным реализациям;

— единовременные расщепления и возвраты состояний с последующими эпохами „спокойного“ обучения.

Алгоритм функционирует следующим образом. После нескольких первичных итераций обучения одногауссовых моделей создаются резервные копии всех состояний, и для каждого состояния запоминается текущее значение его кроссвалидационного критерия (1). Далее производится расщепление всех состояний и выполняется несколько итераций обучения. В конце эпохи обучения сравниваются текущие значения кроссвалидационных критериев состояний с их предыдущими записанными значениями. Если после эпохи обучения кроссвалидационный критерий состояния уменьшился, то это состояние возвращается назад к его резервной реализации. Для всех остальных состояний с возросшим кроссвалидационным критерием применяются те же операции, что и перед первым расщеплением: резервное копирование текущих реализаций состояний и запись текущих значений кроссвалидационных критериев, при этом предыдущие резервные копии удаляются. Таким образом, посредством использования резервных копий состояний обеспечивается рост кроссвалидационных критериев.

По мере роста числа гауссиан количество состояний, возвращаемых к их предыдущим реализациям, начинает превалировать над числом состояний, для которых необходимо увеличить количество гауссиан в смеси. В конце обучения, когда почти все состояния имеют оптимальное количество гауссиан, может возникнуть ситуация, при которой после последнего расщепления рост критериев сменится спадом, поэтому обучение необходимо останавливать непосредственно после откатов состояний к их резервным копиям. Останов обучения производится либо после достижения желаемого количества расщеплений, которое должно быть не

меньше предполагаемого числа гауссиан в состояниях, либо когда прирост ΔC критерия (2) в ходе обучения перестанет превышать заданное значение.

Производить расщепление или возврат к резервной копии каждого состояния в отдельности вместо единовременной процедуры расщеплений и откатов нецелесообразно, поскольку постоянные скачкообразные изменения состояний крайне замедляют процедуру обучения в целом.

Результаты. Эксперименты проводились на базе русской речи SpeechDat(E) [8], содержащей телефонные записи фонетически сбалансированных предложений, слов, словосочетаний, чисел и числовых последовательностей. Процедура сравнения эффективности традиционного обучения с подходом, предложенным в настоящей работе, разбита на две группы экспериментов: на полной базе (около 67 часов речи) и 15 %-ной случайной выборке с целью моделирования недостатка данных (примерно 10 часов речи), который в случае применения метода обучения без контроля состояний быстро приводит к эффекту переобучения и значительно ухудшает качество распознавания посредством обученных подобным образом акустических моделей. Связывание состояний трифонных моделей проводилось с помощью дерева решений [1]. Всего имелось 18 800 связанных состояний для моделей, обученных по полной базе, и около 10 000 связанных состояний для моделей, обученных по 15 %-ной выборке.

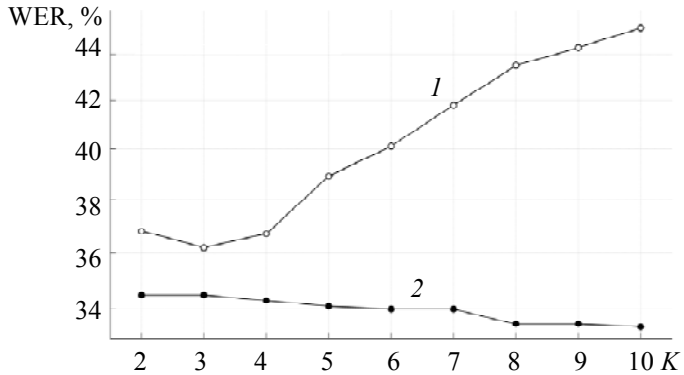
Тестирование качества распознавания речи велось по словарю в 12 500 слов без использования языковой модели. Параметры системы распознавания подбирались таким образом, чтобы обеспечить одинаковую скорость распознавания акустическими моделями, обученными обоими сравниваемыми методами. Результаты обучения (пословная ошибка WER и среднее число гуссиан $\overline{N}_g$) на полном объеме данных на 15 %-ной выборке приведены в таблице.

Результаты обучения

Метод обучения	Полная база		15 %-ная выборка	
	WER, %	$\overline{N}_g$	WER, %	$\overline{N}_g$
Традиционный	34,3	11	45,2	11
Предлагаемый	32,2	3,2	33,4	2,1

Из таблицы видно, что по сравнению с методом обучения без кроссвалидационного контроля предложенный метод дает ощутимое преимущество, более отчетливо проявляющееся в случае недостаточного количества данных для обучения при значительно меньшем количестве гауссиан в состояниях.

На рисунке приведены зависимости ошибки обучения на 15 %-ной выборке от числа K произведенных расщеплений гауссовых смесей. Отчетливо видно, что без кроссвалидационного контроля состояний обычный метод обучения (*1*) приводит к ухудшению качества распознавания, что является свидетельством эффекта переобучения, которого не наблюдается при предложенном подходе (*2*).



Отметим, что индивидуальный контроль состояний позволил достичь приемлемого качества обучения HMM-модели даже на небольшой части обучающей базы. Результаты, полученные этими моделями, уступают полученным на полной базе, только 1 % (абсолютных).

Заключение. В настоящей работе предложена методика обучения акустических моделей с последовательным наращиванием количества гауссиан в состояниях HMM-модели, контролируемым на уровне состояний посредством введения критериев, вычисляемых на кроссвалидационной выборке речевой базы данных. Полученные результаты демонстрируют эффективность метода с точки зрения повышения качества обученных HMM одовременно с сокращением сложности самих моделей вследствие оптимального выбора количества гауссиан.

В качестве развития метода следует рассмотреть возможность использования более тонких кроссвалидационных критериев, поскольку очевидно, что предложенный метод их вычисления — это лишь один из возможных способов ввести меру, которая бы позволила судить о необходимости дальнейших расщеплений состояний. В частности, применение вероятностей посещения, возвращаемых алгоритмом „прямого-обратного“ хода, по-видимому, может повысить эффективность метода. Процедура применения резервных копий состояний для возвратов в случае необходимости к их предыдущим реализациям также может быть модифицирована. Использование нескольких резервных копий состояний в совокупности с алгоритмами отбора этих копий также может повысить качество обучения.

Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

СПИСОК ЛИТЕРАТУРЫ

1. *Young S., Evermann G., Hain, T., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P.* The HTK Book. Cambridge University Engineering Dept. [Электронный ресурс]: <http://htk.eng.cam.ac.uk/docs/docs.shtml>.
2. *Schwarz P.* Phoneme recognition based on long temporal context. PhD thesis. – Brno University of Technology, 2008 [Электронный ресурс]: <http://www.fit.vutbr.cz/~schwarzp/publi/thesis.pdf>.
3. *Andrew J. N.* Model Reduction via the Karhunen-Loeve Expansion. Part I: An Exposition. Technical Report. University of Maryland [Электронный ресурс]: <http://drum.lib.umd.edu/bitstream/1903/5751/1/TR_96-32.pdf>.
4. *Huang X., Acero A., Hon H.W.* Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall, 2001. 1008 p.
5. *Khokhlov Y., Kiselev V., Tampel I., Tatarnikova M.* Phone Recognition Driven Method for Creating Context-Dependent Phones // Proc. of the 12th Intern. Conf. on Speech and Computer (SPECOM 2007). Moscow, 2007. P. 499—502.
6. *Povey D.* Discriminative training for large vocabulary speech recognition. PhD thesis. – Cambridge University Engineering Dept, 2003 [Электронный ресурс]: <https://sites.google.com/site/dpovey/phd_2003.pdf?attredirects=0>.
7. *Khokhlov Y, Tomashenko N.* Speech Recognition Performance Evaluation for LVCSR System // Proc. of the 14th Intern. Conf. on Speech and Computer (SPECOM 2011). Kazan, 2011. P. 129—135.
8. *Heuvel V. H., Boudy J., Bakcsi Z., Cernocky J., Galunov V., Kochanina J., Majewski W., Pollak, P, Rusko M., Sadowski J., Staroniewicz P., Tropf H.S.* SpeechDat-E: Five Eastern European Speech Databases for Voice-Operated Teleservices Completed // Proc. of 7th Europ. Conf. on Speech Communication and Technology (Interspeech 2001). Scandinavia, 2001. P. 2059—2062.

Сведения об авторах

Герман Анатольевич Черных — канд. физ.-мат. наук; ООО „ЦРТ“, Санкт-Петербург; научный сотрудник; Санкт-Петербургский государственный университет; кафедра проблем конвергенции естественных и гуманитарных наук; доцент; E-mail: chernykh@speechpro.com

Максим Львович Кореневский — канд. физ.-мат. наук; ООО „ЦРТ-Инновации“, Санкт-Петербург; научный сотрудник; E-mail: korenevsky@speechpro.com

Кирилл Евгеньевич Левин — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; руководитель отдела распознавания речи; E-mail: levin@speechpro.com

Ирина Александровна Пономарева — ООО „ЦРТ“, Санкт-Петербург; научный сотрудник; E-mail: ponomareva@speechpro.com

Наталья Александровна Томашенко — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ“, Санкт-Петербург; младший научный сотрудник; E-mail: tomashenko-n@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 81'322.6

П. Г. Чистиков, О. Г. Хомицевич, С. В. Рыбин

СТАТИСТИЧЕСКИЕ МЕТОДЫ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ МЕСТ И ДЛИТЕЛЬНОСТИ ПАУЗ В СИСТЕМАХ СИНТЕЗА РЕЧИ

Рассмотрены статистические методы определения местоположения и длительности пауз в системе синтеза речи. Применение таких методов позволяет добиться лучших результатов по сравнению с использованием алгоритмов, основанных на правилах.

Ключевые слова: *пауза, синтез речи, статистические модели.*

Введение. Корректная просодическая разметка в системах синтеза речи необходима для естественного звучания синтезированной речи. Обычно достаточно длинные предложения разбиваются на отдельные фрагменты, которые разделяются паузами. Такие паузы делают речь более понятной и естественной, разрешая неоднозначные трактовки смысла предложений.

Многие системы синтеза речи при определении мест пауз опираются только на знаки препинания. Однако большие участки текста, расположенные между этими знаками, могут звучать монотонно и осложнять восприятие речи, что делает актуальной задачу определения мест пауз на подобных участках. При синтезе русской речи дополнительно возникает другая проблема — знаки пунктуации традиционно используются для обособления различных вводных конструкций, таких как „может быть“, „конечно“ и т.д., которые не выделяются паузами в устной речи.

Кроме того, системы синтеза речи должны не только определять места пауз, но и их продолжительность как внутри предложений, так и между ними. Самым простым решением этой задачи является задание различных констант, регламентирующих длительность пауз. Но так как длительность естественных (в речи человека) пауз является очень вариативной величиной, необходим специальный метод, позволяющий вычислять длительность пауз в зависимости от контекста и структуры предложения [1].

Использование пауз в естественной речи зависит от ряда факторов. Наиболее значимым из них является синтаксическая структура предложения: паузы зачастую располагаются между синтаксически связными компонентами [2, 3]. Однако длина предложения, семантика определенных слов и другие особенности также имеют значение [4]. В системах синтеза речи эти факторы могут быть учтены путем задания правил, определяющих, после какого слова в предложении должна стоять пауза [5, 6], или путем обучения статистических моделей на большом речевом корпусе, на основе которых будут вычисляться вероятности наличия пауз после того или иного слова [7, 8].