7. *Keogh E., Ratanamahatana C.A.* Exact indexing of dynamic time warping // Knowledge and Information Systems. 2005. Vol. 7, Is. 3. P. 358—386.

8. *Алейник С. В., Матвеев Ю. Н., Раев А. Н.* Метод оценки уровня клиппирования речевого сигнала // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 3 (79). С. 79—83.

Сведения об авторах

Виталий Владимирович Киселёв — ООО „Речевые технологии“, Минск; директор; E-mail: kiselev-v@speechpro.com

Андрей Владимирович Ткаченя — ООО „Речевые технологии“, Минск; младший научный сотрудник; E-mail: tkachenia-a@speechpro.com

Михаил Васильевич Хитров — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; генеральный директор; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; заведующий кафедрой; E-mail: khitrov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 004.934

Н. А. Томашенко, Ю. Ю. Хохлов

ИССЛЕДОВАНИЕ ПРОБЛЕМЫ СБАЛАНСИРОВАННОСТИ ДАННЫХ ПРИ ПОСТРОЕНИИ АКУСТИЧЕСКИХ МОДЕЛЕЙ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Исследована проблема сбалансированности данных при обучении акустических моделей для автоматического распознавания речи. Предложена метрика, позволяющая при кластеризации состояний трифонов явно учитывать влияние количества данных в кластере. Экспериментально доказано, что использование такого подхода позволяет повысить качество распознавания речи.

Ключевые слова: *автоматическое распознавание речи, GMM-HMM, обучение акустических моделей, связывание состояний, сбалансированность данных, кластеризация, трифоны.*

Введение. Качество системы автоматического распознавания речи в значительной степени определяется характеристиками используемых в ней акустических моделей. В настоящее время в области распознавания речи обычно применяются статистические подходы, при этом свойства акустических моделей во многом зависят от характеристик речевой базы данных, на которой эти модели были обучены. Одна из наиболее распространенных проблем, связанных с речевыми базами данных, — различие объемов (несбалансированность) данных, приходящихся на разные акустические классы, что может оказывать серьезное влияние на классифицирующую способность моделей [1]. В частности, отсутствие необходимого количества данных в обучающей выборке для определенных моделей усложняет получение надежной оценки параметров этих моделей.

Проблеме несбалансированности классов уделено много внимания в литературе по машинному обучению (см., например, [2]). Несмотря на то что многие алгоритмы обучения предполагают сбалансированность данных, это условие не всегда выполняется для реальных приложений, когда одни классы представлены большим количеством данных в обучающей выборке, а другие — всего несколькими элементами. Этой особенностью отличаются и речевые базы данных, используемые при построении акустических моделей.

Методы решения задачи, связанной с несбалансированностью данных, можно разделить на те, которые направлены на модификацию обучающей выборки и ее балансировку [1, 3], и те, которые преобразовывают сам алгоритм обучения (к последним относится метод, предложенный в настоящей работе).

Цель настоящей работы — исследование влияния сбалансированности данных на качество распознавания. Предметом исследования являются скрытые марковские модели (Hidden Markov Models, HMM), в которых состояния моделей фонем или контекстных трифонов (фонем с определенным левым и правым контекстом) описываются с помощью смеси гауссовых распределений (Gaussian Mixture Models, GMM).

Связывание состояний моделей трифонов при обучении акустических моделей является центральным механизмом в регулировании соотношения между сложностью модели (количеством параметров) и количеством данных в обучающей выборке. Редкие трифоны на уровне состояний или моделей [4] связываются с другими трифонами.

Многие алгоритмы обучения акустических моделей (кластеризации трифонов) [4—10] — агломеративный data-driven (управляемый данными), tree-based clustering (кластеризация на основе фонетического дерева) или их модификации — не позволяют напрямую задавать степень влияния количества данных на параметры связывания состояний трифонов. В настоящей работе предложена метрика, позволяющая по количеству данных в процессе кластеризации учитывать сбалансированность классов, а также исследована зависимость качества акустических моделей от степени этого влияния.

Кластеризация трифонов. Существует два основных метода кластеризации состояний трифонов при обучении акустических моделей на основе GMM-HMM.

1. Агломеративная кластеризация. Этот метод представляет собой „восходящую“ (bottom-up) процедуру кластеризации. Изначально все состояния трифонов рассматриваются как отдельные кластеры. Далее выбирается пара кластеров, при объединении образующих наименьший кластер. Процесс объединения кластеров продолжается до тех пор, пока размер самого большего кластера не достигнет заданного порога либо пока число кластеров не станет равным заданному значению. Размер кластера определяется как наибольшее расстояние между двумя входящими в него состояниями. В качестве расстояния может быть использовано взвешенное евклидово расстояние между средними гауссиан (для случая, когда состояния трифонов описываются одногауссовыми распределениями) либо другие виды метрик [6, 9].

2. Кластеризация на основе фонетического дерева. Этот метод [5, 10] использует бинарное фонетическое дерево решений, его преимущество — возможность моделирования трифонов, которых не было в обучающей выборке. Каждой вершине фонетического дерева соответствуют вопросы (с ответом „да/нет“), относящиеся к свойствам трифонов. Изначально все состояния, подлежащие кластеризации, образуют один класс (в вершине дерева). Далее на каждом шаге алгоритма, в зависимости от выбранного вопроса, происходит расщепление листа дерева на 2 класса. Вопрос в каждой вершине выбирается таким образом, чтобы максимизировать (локально) значение функции правдоподобия на обучающей выборке.

Предлагаемый алгоритм кластеризации состояний трифонов основан на стандартном агломеративном подходе с использованием приведенной ниже метрики, позволяющей в процессе связывания состояний явно учитывать количество данных (векторов в обучающей выборке), приходящихся на отдельные кластеры, и регулировать степень влияния распределения данных по кластерам на параметры связывания. В настоящей работе исследуется задача кластеризации данных для построения модели GMM-HMM. Акустическими единицами речи здесь являются контекстные трифоны. Каждый трифон состоит из трех состояний. Состояния трифонов описываются смесями гауссовых распределений.

1. Метрика должна отражать специфику данных рассматриваемой задачи. Сначала введем расстояние между двумя гауссианами. Расстояние от гауссианы G_1 до G_2 определим следующим образом с помощью расстояния Махаланобиса:

$$\rho(G_1, G_2) = \sqrt{\sum_{i=1}^{n} \frac{\left(\mu_{1i} - \mu_{2i}\right)^2}{\left(\sigma_{1i}\right)^2}}. \quad (1)$$

Здесь μ_{ji} — координата i среднего вектора гауссианы G_j, σ_{1i} — элемент диагональной ковариационной матрицы гауссианы G_1, n — размерность вектора признаков.

Для того чтобы расстояние было симметричным, используем следующую метрику:

$$\overline{\rho}(G_1, G_2) = \frac{1}{2}\left(\rho(G_1, G_2) + \rho(G_2, G_1)\right). \quad (2)$$

Процедура кластеризации состояний трифонов с использованием предложенной метрики включает следующие шаги. В начале процесса связывания все состояния трифонов описываются одногауссовыми распределениями. Далее происходит последовательное объединение гауссиан в классы, при этом на каждом шаге объединяются наиболее близкие классы.

Близость классов определим следующим образом: каждый класс C описывается множеством гауссиан его состояний, расстояние между двумя классами C_1 и C_2 вычисляется по формуле:

$$\rho(C_1, C_2) = \frac{1}{\sum\limits_{\substack{k \in C_1 \\ m \in C_2}} \left(N_k + N_m\right)} \sum_{\substack{k \in C_1 \\ m \in C_2}} \left(N_k + N_m\right) \overline{\rho}(G_k, G_m). \quad (3)$$

Здесь N_k — число векторов в обучающей выборке для состояния, которому соответствует гауссиана G_k.

2. Учет количества данных при кластеризации состояний трифонов. Для преодоления проблемы сбалансированности данных предлагается модифицировать метрику расстояний между классами (3) следующим образом:

$$\rho_{\text{bal}}(C_1, C_2) = \rho(C_1, C_2)(N_1 + N_2)^p. \quad (4)$$

Здесь p — степень влияния количества данных на метрику связывания, N_i — число векторов в кластере C_i.

Важно отметить, что оценки статистик (и расстояний) для состояний трифонов, у которых слишком мало данных в обучающей выборке, недостоверны. Поэтому в алгоритме кластеризации были выделены два этапа:

1) *слияние всех достаточно малых классов.* Если при выборе классов для объединения число векторов, приходящихся на какой-либо из них, меньше заданного порога thr, то расстояние между ними умножается на малое положительное число ε (например, $\varepsilon = 10^{-3}$);

2) *кластеризация всех остальных классов.*

Эти два этапа можно совместить, модифицировав метрику расстояний между классами следующим образом:

$$\rho^*(C_1, C_2) = \begin{cases} \rho_{\text{bal}}(C_1, C_2)\varepsilon, & \text{если } \min\left\{N_1, N_2\right\} < \text{thr}, \\ \rho_{\text{bal}}(C_1, C_2), & \text{если } \min\left\{N_1, N_2\right\} \geq \text{thr}. \end{cases} \quad (5)$$

Эксперименты. Цель экспериментов — установить, как влияет учет количества данных при связывании состояний на качество акустических моделей.

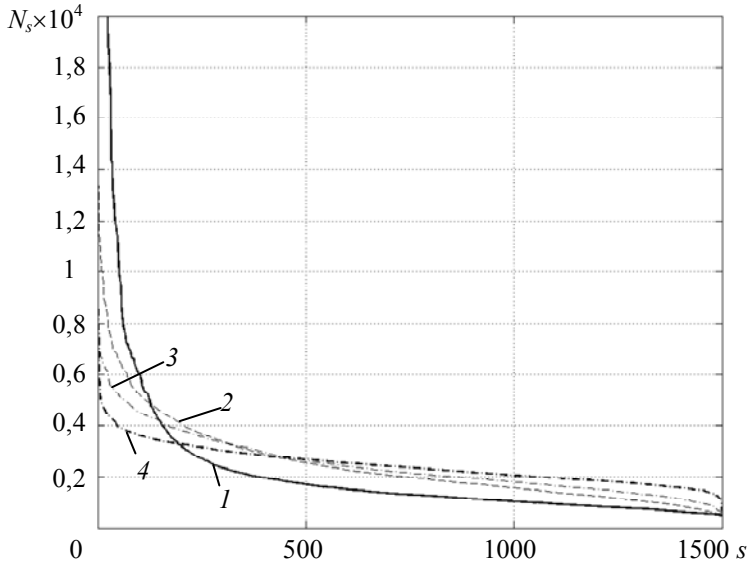
Речевая база данных для обучения состоит из фонограмм русской речи (как чтения, так и спонтанной) более чем 260 дикторов; тип канала — микрофон (16 кГц). Из исходной базы были сформированы четыре обучающие выборки размером 160, 80, 40 и 20 часов речи. Для каждой обучающей выборки были построены четыре акустические модели с разными значениями параметра p (см. (4)): 0; 0,4; 0,7; 2 (p=0 соответствует случаю, когда при кластеризации степень сбалансированности кластеров по количеству данных не учитывается).

Для построения акустических моделей использовались признаки MFCC+Δ+ΔΔ [11]: 12 мел-частотных кепстральных коэффициентов (MFCC, mel-frequency cepstral coefficients) и энергия вместе с производными первого и второго порядка от этих величин по времени. Фонетический алфавит состоит из 52 фонем русского языка и паузы. Во всех экспериментах количество связанных состояний трифонов равно 1500. Параметры обучения акустических моделей (за исключением p) во всех экспериментах совпадают, максимальное число гауссиан в состоянии 25. Тесты распознавания проводились на выборке из 1000 файлов общей длительностью 81 мин. Словарь для распознавания содержит 5400 словоформ. В качестве метрики для оценки качества акустических моделей используется пословная ошибка (Word Error Rate, WER), которая вычисляется следующим образом [12]:

$$\mathrm{WER} = \frac{S + D + I}{N} \cdot 100 .$$

Здесь S, D и I обозначают соответственно число ошибок замен, пропусков и вставок слов при распознавании; N — число слов в тексте, который был произнесен.

На рисунке для одной из обучающих выборок показано распределение количества векторов по состояниям трифонов после кластеризации для разных значений p (*1* — p=0; *2* — 0,4; *3* — 0,7; *4* — 2). На оси абсцисс приведены номера связанных состояний (s) после их упорядочивания в порядке убывания количества векторов N_s, приходящихся на эти состояния после кластеризации.



Результаты экспериментов приведены в таблице. Для каждой выборки показана разница (ΔWER) лучших результатов с учетом сбалансированности данных (p>0) и без ее учета (p=0), $\Delta\mathrm{WER}_{\mathrm{rel}}$ — относительные значения этой разницы. Для p=0,4 и 0,7 в среднем результаты получаются лучше, чем для p=0. Это подтверждает гипотезу о том, что возможность явного влияния на сбалансированность классов при кластеризации позволяет повысить качество аку-

стических моделей. Однако для случая, когда размер базы составляет всего 20 часов, улучшения при $p>0$ не наблюдается.

Результаты распознавания для разных значений p и различного объема обучающих выборок

Размер базы, ч	p	WER	ΔWER	ΔWER_{rel}
160	0	35,4	—	—
	0,4	34,2	—	—
	0,7	33,9	1,5	4,2
	2	34,4	—	—
80	0	35,2	—	—
	0,4	34,0	1,2	3,4
	0,7	34,1	—	—
	2	34,6	—	—
40	0	34,8	—	—
	0,4	33,8	1,0	2,9
	0,7	34,2	—	—
	2	34,4	—	—
20	0	33,8	—	—
	0,4	33,9	—	—
	0,7	33,8	0,0	0,0
	2	34,0	—	—

Заключение. В работе предложен метод учета количества данных при кластеризации состояний трифонов в процедуре обучения GMM-HMM для систем автоматического распознавания речи. Эксперименты показали, что предложенный метод позволяет повысить качество акустических моделей и при правильном выборе степени влияния количества данных позволяет уменьшить WER на 3—4 % относительно исходного значения. Уменьшения WER при учете количества данных удалось достичь только для обучающих выборок объемом не менее 40 часов. Используемая в статье базовая метрика (3) отличается от таких распространенных метрик, как расстояние Евклида, Бхатачария [9] и других [5, 6], но подход (4) можно обобщить на случай других метрик. При использовании базовой метрики для кластеризации состояний без взвешивания с учетом количества данных в классах будут неточными модели, для которых было мало данных в обучающей выборке. Введение метрики, приводящей к сбалансированности классов по количеству обучающих данных, позволяет построить более надежные модели для тех состояний, которые в первом случае оказались плохо обученными, но в то же время может сделать менее точным разделение между остальными моделями. Для лучшей кластеризации необходим выбор параметров метода, обеспечивающий наилучшее соотношение надежности моделей и точности их разделения.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Irtza S., Hussain S.* Minimally balanced corpus for speech recognition // Proc. 1st IEEE Intern. Conf. on Communications, Signal Processing, and their Applications (ICCSPA). 2013. P. 1—6.

2. *Guo X., Yin Y., Dong C., Yang G., Zhou G.* On the class imbalance problem // Proc. IEEE 4th Intern. Conf. on Natural Computation (ICNC'08). 2008. Vol. 4. P. 192—201.

3. *Garcia-Moral A. I., Solera-Ureña R., Peláez-Moreno C., Díaz-de-María F.* Data balancing for efficient training of hybrid ANN/HMM automatic speech recognition systems // IEEE Transact. on Audio, Speech, and Language Processing. 2011. Vol. 19, N 3. P. 468—481.

4. *Darjaa S., Cernak M., Trnka M., Rusko M., Sabo R.* Effective Triphone Mapping for Acoustic Modeling in Speech Recognition // Proc. INTERSPEECH. 2011. P. 1717—1720.

5. *Young S. J., Odell J. J., Woodland P. C.* Tree-based state tying for high accuracy acoustic modelling // Proc. of the Workshop on Human Language Technology. Association for Computational Linguistics. 1994. P. 307—312.

6. The HTK book / *Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Woodland P.* // Cambridge University Engineering Department. 2002.

7. *Aubert X., Beyerlein P., Ullrich M.* A bottom-up approach for handling unseen triphones in large vocabulary continuous speech recognition // Proc. IEEE 4th Intern. Conf. on Spoken Language (ICSLP 96). 1996. Vol. 1. P. 14—17.

8. *Park J., Ko H.* Effective acoustic model clustering via decision-tree with supervised learning // Speech communication. 2005. Vol. 46. N 1. P. 1—13.

9. *Mak B., Barnard E.* Phone clustering using the Bhattacharyya distance // Proc. IEEE 4th Intern. Conf. on Spoken Language (ICSLP 96). 1996. Vol. 4. P. 2005—2008.

10. *Wang G., Sim K. C.* An investigation of tied-mixture GMM based triphone state clustering // Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). 2012. P. 4717—4720.

11. *Матвеев Ю. Н.* Исследование информативности признаков речи для систем автоматической идентификации дикторов // Изв. вузов. Приборостроение, 2013. Т. 56, № 2. С. 47—51.

12. *Khokhlov Y., Tomashenko N.* Speech Recognition Performance Evaluation for LVCSR System // Proc. 14th Intern. Conf. “SPEECH and COMPUTER” (SPECOM 2011). Kazan. Russia. 2011. P. 129—135.

Сведения об авторах

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

Юрий Юрьевич Хохлов — ООО „ЦРТ“, Санкт-Петербург; ведущий программист;
E-mail: khokhlov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.