СИСТЕМА РАСПОЗНАВАНИЯ СЛИТНОЙ РЕЧИ НА ОСНОВЕ МАШИН НА ОПОРНЫХ ВЕКТОРАХ

И. Э. Хейдоров*, А. А. Кузьмин**

Белорусский государственный университет,
факультет радиофизики и компьютерных технологий
Минск, Беларусь
*E-mail: igorhmm@mail.ru**
*E-mail: kuzAleksAleks@gmail.com***

Предложена система распознавания слитной речи, основанная на машинах на опорных векторах (МОС), а также представлены результаты сравнения точности распознавания разработанной системы по сравнению с классической схемой, использующей связку скрытых марковских моделей (СММ) и гауссовой смеси. Алгоритм работы системы включает два основных шага: сначала при помощи МОВ акустический сигнал преобразуется в зашумленную последовательность символов. На втором этапе, методы сравнения строк используются для определения последовательности слов, соответствующих произнесенной фразе.

Ключевые слова: распознавание слитной речи, машины на опорных векторах, скрытые марковские модели, гауссова смесь, сравнение строк.

Введение

Несмотря на быстрое развитие новых методов распознавания образов, связка СММ и гауссовых смесей стала стандартом для систем распознавания слитной речи.

Скрытая марковская модель является стохастическим конечным автоматом, состоящим из конечного множества состояний $Q = \{q_1, \dots, q_{1K}\}$ и вероятностей непосредственных переходов между ними. Каждое такое состояние связано с соответствующей эмиссионной плотностью распределения вероятности $p(x_n|q_k)$. Таким образом, СММ может быть использована для моделирования последовательности векторов признаков X как кусочно-стационарного процесса, в котором каждый стационарный участок относится к определенному состоянию СММ. Такой подход обеспечивает моделирование динамической структуры речевой единицы (т. е. решает проблему различной длительности сигнала, соответствующего одной и той же фонеме), а также дискриминационную способность на локальных участках речевого сигнала [1].

Как правило, в системах, основанных на СММ, классификацию участка сигнала, относящегося к тому или иному состоянию СММ, осуществляет гауссова смесь, которая по своей дискриминационной и обобщающей способности уступает машинам на опорных векторах (МОВ). Поэтому МОВ был с успехом интегрирован в системы распознавания речи, основанные на СММ, заменив в них гауссову смесь для моделирования плотности распределения эмиссионной вероятности [1, 3, 4].

Машины на опорных векторах [2, 5] − относительно новый метод в сфере распознавания речи, который обладает следующими преимуществами:

1. Хорошая обобщающая способность даже при относительном малом количестве обучающих данных;
2. Алгоритм позволяет найти уникальную разделяющую гиперплоскость для классификации, за счет сходимости к глобальному минимуму целевой функции при обучении.

Одним из главных недостатков МОВ можно считать время обучения, которое имеет как минимум квадратичную [6] зависимость от количества обучающих векторов. При использовании машин на опорных векторах в связке со скрытыми марковскими моделями эта проблема становится особенно значимой в случае большого количества обучающих данных.

Основной идеей данной работы является отдельное использование мультиклассовых МОВ для распознавания монофонов в слитной речи.

Результатом распознавания фонем вдоль сигнала становится зашумленная транскрипция произнесенной фразы. Задача заключается в нахождении слов, транскрипция которых в наибольшей степени соответствует полученной последовательности фонем. Эту проблему можно свести к проблеме поиска общих подстрок в двух строках. Такая задача решается уже довольно давно [9, 10]. В данной работе для определения общих подстрок и расчета коэффициента подобия применяется алгоритм Майерса [11].

Таким образом, динамическая последовательность фонем моделируется с помощью методов сравнения строк. Кроме того, представлено расширение алгоритма для распознавания фраз, состоящих из нескольких слов.

Распознавание фонем с помощью мультиклассовой машины на опорных векторах

Первоначально машины на опорных векторах представляют собой особый вид классификатора, который находит соответствие между вектором данных x_l и одним из двух классов y_l $\{-1, +1\}$ в соответствии со знаком выражения:

$$y = \sum_{l=1}^{L} \lambda_l y_l K(x_l, x) - b, \quad (1)$$

где $K(\cdot,\ \cdot)$ – симметричное положительно-определенное ядро интегрального уравнения, которое подчиняется ряду ограничений [7], поскольку является скалярным произведением в определенном пространстве:

$$K(x_l, x) = \phi(x_l)^T \phi(x), \quad (2)$$

где ϕ – это функция, переводящая вектор исходных данных в пространство с большей (возможно бесконечной) размерностью.

В качестве ядра часто выступают полиномиальные функции, радиальная базисная функция и сигмаидальная функция. Параметры λ_l и b определяются исходя из решения задачи квадратичного программирования с линейными ограничениями [8]. Основное преимущество метода МОВ основывается на том, что только малая часть коэффициентов λ_l отлична от нуля. Другими словами, только малый набор опорных векторов (x_l с ненулевыми λ_l) нужен для классификации согласно выражению (1).

При построении системы распознавания речи общее количество фонем составило 42. Поэтому использовался мульти-классовый классификатор из 861 классификатора (по одному на каждую пару классов).

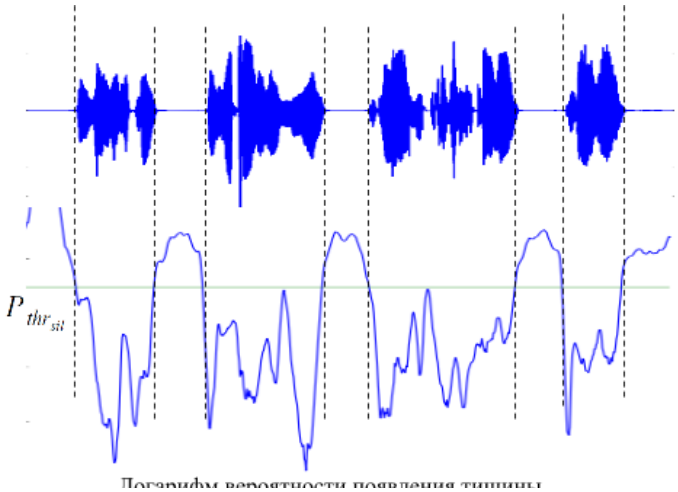

Рис. 1. Выделение участков, относящихся к произнесенным словам на фоне тишины с помощью акустической модели на основе связки СММ и гауссовых смесей

Стоит отметить, что для выделения произнесенной фразы на фоне тишины использовалась акустическая модель на основе связки СММ и гауссовой смеси следующим образом: модель применяется для расчета вероятности появления тишины на каждом временном шаге вдоль всего сигнала. С помощью заранее установленного порога $P_{thr_{sil}}$ определяются участки сигнала, которые с наименьшей вероятностью соответствуют тишине (рис. 1). Именно они и рассматриваются далее в распознавании:

$$P\left(X\left(t,T_{sil}\right)\middle|\Lambda_{sil}\right) < P_{thr_{sil}}, \quad X\left(t,T_{sil}\right) = x_t, x_{t+1}, \ldots, x_{t+T_{sil}} \tag{3}$$

где $X\left(t, T_{sil}\right)$ – последовательность векторов наблюдений, начиная с момента времени t, T_{sil} – длина последовательности, Λ_{sil} – параметры акустической модели.

Определение слов по зашумленной транскрипции

Как уже было сказано выше, после распознавания фонем основной задачей является сравнить полученную зашумленную строку фонем с транскрипциями из словаря для определения наиболее подходящей последовательности слов (рис. 2).

После распознавания фонем последовательность включает в себя длинные подпоследовательности подряд идущих одинаковых символов. Поэтому требуется начальная обработка для сокращения подобных участков до одной соответствующей фонемы.

Транскрипция каждого слова в словаре сравнивается с зашумленной последовательностью фонем согласно алгоритму Майерса [11]. Те слова, транскрипция которых имеет наибольшее совпадение, принимаются как результат распознавания. Каждое последующее слово сравнивается с той частью зашумленной последовательности фонем, которая следует за уже распознанной частью.

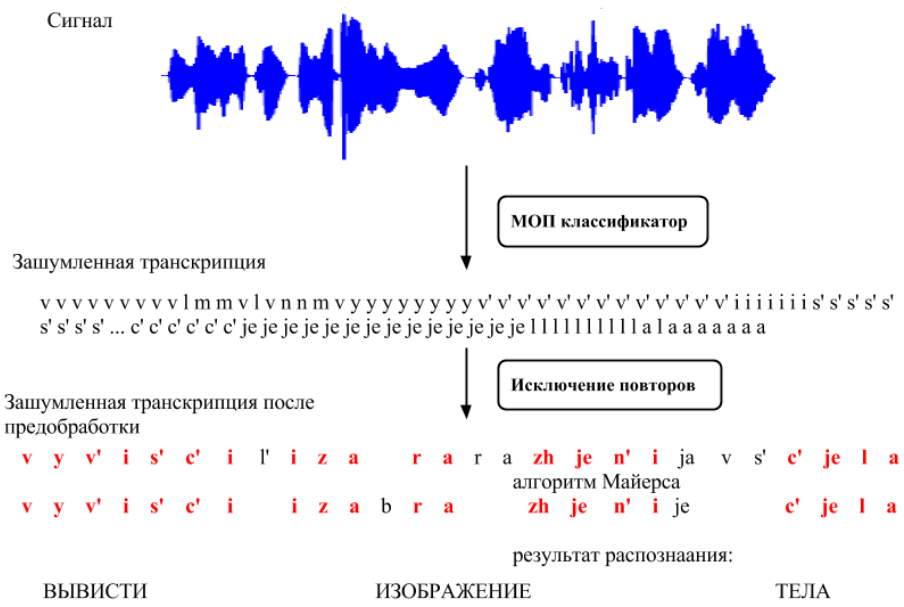

Рис. 2. Диаграмма алгоритма распознавания

Результаты эксперимента и выводы

Для обучения классификатора на основе МОВ использовался корпус, состоящий из записей русской и белорусской речи. В записи участвовал один диктор. Общая продолжительность звучания всех записей корпуса − 3397 секунд. Так как для МОВ классификатора используется «обучение с учителем», все записи размечены на монофоны вручную. Словарь включает в себя 56 слов.

В тестировании использовалась база данных записей русской речи общей продолжительностью звучания 840 секунд. В записи, как и в случае с обучающей базой, участвовал один диктор.

Перед обучением и тестированием сигнал преобразовывался в последовательность векторов признаков, представляющих из себя 13 кепстральных коэффициентов, распределенных по шкале МЭЛ-частот. В состав векторов были включены приближения первой и второй производной каждого коэффициента. Общая размерность пространства векторов признаков составила, таким образом, 39.

Сравнение проводилось с базовой системой распознавания на основе связки СММ и гауссовых смесей. Система моделировала каждую фонему с помощью СММ из 3-х эмитирующих состояний. Гауссова смесь для каждого состояния состояла из одной нормальной функции с диагональной матрицей ковариации.

Результаты сравнения точности распознавания систем представлены в табл. 1.

Таблица 1

Сравнение точности распознавания для различных систем

	Точность распознавания слов, %	Точность распознавания предложений, %
Система на основе МОВ, ядро: полином степени 3	81,44	38,03
Система на основе МОВ, ядро: радиальная базисная функция	90,38	**67,61**
Система на основе связки СММ и гауссовой функции	**92,47**	60,50

Таким образом, разработанная система позволила получить точность распознавания предложений, превосходящую ту же точность для классической системы на основе связки СММ и гауссовой смеси при прочих равных условиях.

Стоит, однако, отметить, что в данном случае не накладывалось ограничений на последовательности слов. Кроме того, точность определялась для записей, сделанных одним диктором, что позволило использовать обобщающую способность МОВ.

В ходе дальнейших исследований будет определяться точность распознавания предложенной системы на записях большего объема, сделанных несколькими дикторами.

Библиографические ссылки

1. *Kruger S. E., Schaffoner M., Katz M., Andelic E., Wendemuth A.* Mixture of Support Vector Machines for HMM based Speech Recognition // The 18th International Conference of Pattern Recognition ICPR'06, 2006 IEEE.
2. *Vapnik C. N.* Statistical Learning Theory // A Wiley-Interscience Publication, 1998.
3. *Golowich S. E., Sun D. X.* A support vector/hidden markov model approach // In ASA Proc. Of the Statistical Computing Section. 1998. P. 125–130.
4. *Stadermann J., Rigoll G.* Ahybrid SVM/HMM acoustic modeling approach to automatic speech recognition // In Proc. Of INTERSPEECH-2004 ICSLP. 2004. P. 661–664.
5. *Vapnik V. N., Cortes C.* Support-VectorNetworks // Machine Learning, 1995. 20. P. 273–297.
6. *Collobert R., Bengio S., Bengio Y.* Parallel mixture of svms for very large scale problems // In ASA Proc. Of the Statistical Computing Section, 2002. 14. P. 1105–1114.
7. *Boser B., Guyon I., Vapnik V.* A training algorithm for optimal margin classifier // in Proc. Of the Fifth Annual ACM Workshop on Computational Learning Theory. 1992. P. 144–152.
8. *Vapnik V.*The Nature of Statistical Learning Theory / V. Vapnik, Springer, Second Edition, 1999.
9. *Aho A. V., Hirschberg D. S., Ullman J. D.* Bounds on the Complexity of the Longest Common Subsequence Problem // Journal of ACM 23. 1. 1976. P. 1–12.
10. *Hirschberg D. S.* Algorithms for the Longest Common Subsequence Problem // Journal of ACM 24. 4. 1977. P. 664–675.
11. *Myers E. W.* An O(ND) Difference Algorithm and Its Variations // Algorithmica1(2) 1986. P. 251–266.