А. В. Максимов, **С. Ю. Мельников**, **Н. М. Чавчавадзе** (Москва, ТВП; ООО «Стэл-КС»). **Тенденции развития методов автоматической идентификации языка речевых и текстовых сообщений.**

Введение. Актуальность тематики идентификации языка текстовых и речевых сообщений в последние годы возрастает. Количество публикаций, посвященных этой задаче, начало интенсивно расти с середины 90-х годов и сейчас составляет пять–шесть десятков в год. Анализ наиболее современных публикаций позволяет выделить следующие основные тенденции:

— смещение акцентов с «наивно-лингвистических» подходов на более точные методы идентификации, основанные на статистической теории распознавания образов;

— изучение все более сложных и точных вероятностных моделей языка, включающих семантические аспекты;

— появление в последние годы сверхскоростных аппаратных платформ, применение которых позволяет реализовать средства определения языка на новом качественном уровне;

— уточнение и детализация формулировок задачи идентификации языка: для текстов определение не только основного языка, но второго и других, для речи определение акцента и диалекта.

Далее на основании анализа материалов нескольких международных конференций приведены основные характеристики и особенности современных методов, применяемых для решения указанной задачи, формулируется ряд выводов и практических предложений.

Идентификация языка текста. Наибольшую эффективность в настоящее время дают следующие методы: словарный, метод частых в данном языке слов, метод уникальных сочетаний букв, марковские n-граммные модели [1, 2]. Точность идентификации языка, достигаемая этими методами, колеблется в пределах 98-99,9%. Максимальная точность может достигаться, например, с применением марковских n-граммных моделей (обычно используются $n = 3, 4$), обученных на больших выборках, с использованием различных методов сглаживания вероятностей n-грамм больших порядков [8].

Трудоемкость таких методов имеет линейную зависимость от длины текста. Одним из наиболее трудоемких является метод марковских моделей, кроме того в данном методе нужно хранить большое количество данных о моделях языка. Этот метод используется в большинстве программных решений для идентификации языка текста. Для скоростных аппаратных решений в [3] предложена модификация метода, использующая «усеченную» n-граммную модель. При подсчете статистики критерия используются не все n-граммы в тексте, а только те, которые содержатся внутри слов и обладают наибольшей различающей способностью заданного набора языков. Кроме того, в целях экономии памяти, используется редукция алфавита текста к латинице, что позволяет каждую букву кодировать пятью битами. Предложенная модификация реализована с помощью специальной платы-ускорителя с ПЛИС Xilinx [3]. Аппаратное решение для этой задачи на основе ПЛИС Altera описано в [4]. Достигаемые с помощью указанных ускорителей скорости обработки текста превышают 2 Гб/с.

Идентификация языка речевого сообщения. Задача определения языка речевого сообщения, несмотря на кажущуюся близость к задаче определения языка текста, является по сравнению с ней значительно более сложной и вычислительно трудоемкой.

Неплохие результаты в решении такой задачи в настоящее время получены на основе акустического подхода, использующего аппарат гауссовских смесей (GMM). Более высокую точность обеспечивает фонотактический подход, связанный с использованием тех или иных фонетических распознавателей (фонов, слогов, артикуляционных признаков и пр.), и последующим анализом статистических свойств последовательности распознанных элементов. Лучшие результаты по идентификации языков

получены на основе параллельного использования большого арсенала различных методов [5], [6]. Использование нескольких систем признаков приводит к необходимости построения конечного классификатора, который отвечает за принятие согласованного решения. Таким классификатором является, как правило, нейронная сеть [7] или гауссовский конечный классификатор.

Точность, достигаемая современными системами, составляет до 97 %. Эта величина зависит от набора идентифицируемых языков, длительности и качества речевого сигнала. Однако требования, которые такие системы предъявляют к вычислительным ресурсам, весьма высоки. Время обработки фрагмента речи на стандартном однопроцессорном компьютере сопоставимо с его длительностью.

Выводы. В последнее время разработаны теоретические подходы для решения задачи идентификации языков, которые обеспечивают приемлемую для практики точность. Для идентификации языков текстов уже существуют достаточно компактные аппаратные средства, позволяющие работать со скоростями до нескольких гигабит в секунду. Для идентификации языка речевого сообщения таких скоростных и надежных методов пока нет, а необходимых на практике скоростей обработки можно достигать путем распараллеливания вычислений и создания высокопроизводительных специализированных ускорителей на базе современных ПЛИС.

СПИСОК ЛИТЕРАТУРЫ

1. *Cavnar W. B.*, *Trenkle J. M.* N-gram-based text categorization. — In: 1994 Symposium on Document Analysis and Information Retrieval in Las Vegas, 1994, p. 161–175.
2. *Dunning T.* Statistical Identification of Language. Technical Report. Las Cruces: Computing Research Laboratory, New Mexico State University. 1994, p. 94–273.
3. *Kastner C. M.*, *Covington G. A.*, *Levine A. A.*, *Lockwood J. W.* HAIL: a hardware-accelerated algorithm for language identification. — 15th Annual Conference on Field Programmable Logic and Applications (FPL), (Tampere, Finland), 24–26 Aug. 2005, p. 499–504.
4. *Jacob A.*, *Gokhale M.* Language classification using n-grams accelerated by FPGA-based Bloom filters. — In: Proceedings of the 1st International Workshop on High-performance Reconfigurable Computing Technology and Applications (in conjunction with SC07), Reno, Nevada. 2007, p. 31–37.
5. *Campbell W.*, *Gleason T.*, *Navratil J.*, *Reynolds D.*, *Shen W.*, *Singer E.*, *Torres-Carrasquillo P.* Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation. — In: IEEE Odyssey 2006: The Speaker and Language Recognition Workshop. San Juan, Puerto Rico, June 2006.
6. *Kwan C.*, *Yin J.*, *Ayhan B.*, *Chu S.*, *Liu X.*, *Pukket K.*, *Zhao Y.*, *Ho K.*, *Kruger M.*, *Sityuar I.* An integrated Approach to Robust Speaker Identification and Speech Recognition. — In: IEEE World Congress on Computation Intelligence. Hong Kong, June 2008.
7. *Joshi S.*, *Joshi S.*, *Prahallad K.*, *Yegnanarayana B.* AANN-HMM Models for Speaker Verification and Speech Recognition. — In: IEEE World Congress on Computation Intelligence, Hong Kong. June 2008.
8. *Stanley F. Chen*, *Joshua Goodman.* An Empirical Study of Smoothing Techniques for Language Modeling. Cambridge, MA: Harvard Univ., Computer Science Group, 1998.