Система автоматического распознавания языков на основе гауссовских и авторегрессионных моделей

Ю.С. Иващенко

Д.А. Леднов,
кандидат технических наук

Н.А. Любимов

Под системой автоматической идентификации языков (Automatic Language Identification — LID) подразумевается такая система, на вход которой поступают записи речевых сообщений, а на выходе формируется заключение о языке сообщения.

Необходимые требования к подобной системе — текстонезависимость и дикторонезависимость. Текстонезависимостью является устойчивость работы системы по отношению к изменению содержания входного сообщения. Под дикторонезависимостью устройства подразумевается способность системы распознавать язык сообщения, сказанного произвольным диктором.

В основе системы автоматической идентификации языков (САИЯ) лежит создание набора различных языковых моделей. Формирование моделей производится на основе фонетико-лингвистических [2] или акустических параметров речи [1]. Первый подход достаточно эффективен в распознавании, но требует много вычислительных ресурсов, что делает его неприменимым для большинства систем реального времени. В настоящей статье рассматривается второй подход, использующий акустическую параметризацию речи.

Цель работы — представить сравнительный анализ трёх различных типов акустических моделей языка.

Алгоритмы предобработки

В работе используются два различных типа предварительной обработки речевых данных. Векторы наблюдения были сформированы на основе:

- Логарифмически масштабированных кепстральных коэффициентов *(Mel Frequency Cepstral Coefficients — MFCC).*
- Смещённого дельта кепстра *(Shifted Delta Cepstrum — SDC) [5]*

Частота дискретизации звуковых данных в обоих случаях составляла 8 кГц с разрядностью 16 бит. Для вычисления коэффициентов MFCC длительность окна анализа выбиралась равной 16 мс, окна анализа следовали друг за другом с шагом 8 мс. Размерность результирующего вектора наблюдений составляла 12.

Длительность окна для вычисления вектора SDC такая же, как и у MFCC. Вектора получены путём конкатенации семи компонент, каждая из которых представляет из себя разность сдвинутых во времени кепстральных коэффициентов для смежных блоков [5].

Вероятностные модели

Полученный в результате предварительной обработки речевых данных набор векторов наблюдения полагается выборкой некоторой генеральной совокупности независимых случайных величин. Плотность распределения этих величин можно описать смесью однотипных распределений. Задача построения вероятностной модели языка сводится к задаче разделения смеси, т.е. оценки параметров распределения каждой из компонент смеси при условии, что общее число компонент и их вид заранее известны.

Рассмотрим два типа вероятностных моделей — смесь нормальных распределений и смесь нормальных распределений авторегрессии. В качестве обобщённого подхода к оценке неизвестных параметров применялся стандартный EM — алгоритм [3]. С помощью максимизации функционала (логарифма правдоподобия):

$$p(\theta \mid x) = \ln \prod_X p(x \mid \theta) = \sum_X \ln\, p(x \mid \theta) \to \max$$

настраиваются параметры каждой из компонент смеси. Оптимизация проводится итерационно методом покоординатного спуска.

Для смеси нормальных распределений базовым объектом является гауссиан-функция, описывающая плотность распределения случайной величины для нормального закона. Метод разделения такой смеси не предполагает дополнительных знаний о способе представления речевых данных, поэтому вероятностная модель подобного типа была построена как для MFCC, так и для SDC. Настраиваемыми параметрами каждой из компонент являются: её априорная вероятность, математическое ожидание и дисперсия.

Смесь нормальных распределений авторегрессии имеет более сложную структуру, позволяющую учитывать временну̀ю зависимость входных данных. Базовой единицей этой модели является функция вида:

$$N_{AR}(x_t \mid \alpha_1,\ldots,\alpha_K,\eta,\Lambda) = \frac{1}{\sqrt{(2\pi)^d \mid \Lambda \mid}} \exp\{-\frac{1}{2}(x_t - \sum_{k=1}^{K}\alpha_k x_{t-k} - \eta)^T \Lambda^{-1}(x_t - \sum_{k=1}^{K}\alpha_k x_{t-k} - \eta)\}, \quad (1)$$

где

$x_t \in R^d$ — переменная, наблюдаемая в момент времени t;

$\alpha_1,\ldots,\alpha_K \in R^d$ — коэффициенты авторегрессии, K — глубина авторегрессии;

$\eta \in R^d$ — постоянная составляющая;

$\Lambda \in R^{d\times d}$ — дисперсия шума (матрица ковариации).

Заметим, что если записать авторегрессионую модель дискретного сигнала:

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_K x_{t-K}, \quad (2)$$

где $\varepsilon_t \sim N(\eta, \Lambda)$ — ошибка линейного предсказания (или ошибка авторегрессии), то функция вида (1) есть плотность распределения ошибки линейного предсказания с математическим ожиданием и ковариационной матрицей Λ[1].

Подобно тому, как это сделано для стандартного EM-алгоритма [3], рассмотрим смесь таких плотностей и зададимся вопросом об оценке параметров смеси.

Смесью нормальных распределений авторегрессии называется функция вида:

$$N(x_t \mid \theta) = \sum_{j=1}^{M} \beta_j N_{AR}^j(x_t, \quad (3)$$

причём $\sum_{j=1}^{M} \beta_j = 1$,

а $\theta^j = (\alpha_1^j, \ldots, \alpha_K^j, \eta^j, \Lambda^j)$ — набор параметров каждой из компонент смеси.

Для того чтобы оценить все неизвестные параметры в выражении (3), воспользуемся методом максимума правдоподобия. Как показано в [3], при помощи искусственного введения вспомогательного вектора «скрытых» переменных функция правдоподобия может быть приведена к виду:

$$Q(\Theta) = \sum_{j=1}^{M} \sum_{i=1}^{N} g_{ij} \log(\beta_j) + \sum_{j=1}^{M} \sum_{i=1}^{N} g_{ij} \log(N_{AR}^j(x_i)) \quad (4)$$

здесь $\Theta = (\beta_1, \ldots, \beta_M, \theta^1, \ldots, \theta^M)$, а g_{ij} — значение «скрытых» переменных, которые определяют, с какой вероятностью i-ый объект был сгенерирован именно j-ой компонентой. Имея некоторые априорные знания о начальном распределении параметров в смеси, эти значения несложно найти, используя формулу Байеса.

Первое слагаемое в выражении (4) не зависит от типа распределения каждой компоненты в смеси. Используя правило множителей Лагранжа и решая задачу условной оптимизации с условием $\sum_j \beta_j = 1$, найдём, что

$$\beta_{onm}^j = \frac{\sum_{i=1}^{N} g_{ij}}{N} \quad (5)$$

Второе слагаемое функции правдоподобия в выражении (4) является квадратичным функционалом по параметрам $\alpha_1^j, \ldots, \alpha_K^j, \eta^j$, $j = 1, \ldots, M$, поэтому необходимым и достаточным условием экстремальности является равенство нулю соответствующих производных. Покажем, что нахождение оптимальных параметров в этом случае эквивалентно нахождению решения системы линейных алгебраических уравнений.

[1] На практике вместо ковариационной матрицы используется вектор дисперсии в предположении о некоррелируемости случайных величин, что практически не ухудшает результат, но значительно упрощает вычисления.

Для любого $s = 1,\ldots,K$ имеем, что

$$\frac{\partial Q}{\partial \alpha_s^j} = \sum_{i=1}^{N} g_{ij}\left[-2\frac{1}{2}(-x_{i-s}^T)\Lambda_j^{-1}(x_i - \sum_{k=1}^{K}\alpha_k^j x_{i-k} - \eta^j)\right] = 0$$ — условие экстремальности.

Далее, ввиду того, что оптимизация параметров каждой из компонент может проходить независимо друг от друга, опустим индекс j.

Из предыдущего выражения следует:

$$\sum_{i=1}^{N} g_i x_{i-s}^T \eta + \sum_{k=1}^{K}\alpha_k \sum_{i=1}^{N} g_i x_{i-s}^T x_{i-k} = \sum_{i=1}^{N} g_i x_{i-s}^T x_i \quad s = 1,\ldots,K \qquad (6)$$

Это система из K уравнений с K+1 неизвестной. Как было отмечено выше, параметр η есть математическое ожидание ошибки линейного предсказания, которая распределена нормально. Поэтому, используя представление из [3], оптимальное значение параметра η имеет вид:

$$\eta = \frac{\sum_{t=1}^{N} g_t (x_t - \sum_{k=1}^{K}\alpha_k x_{t-k})}{\sum_{t=1}^{N} g_t} \qquad (7)$$

Подставляя указанное выше выражение в (6), получим

$$\frac{\sum_{i=1}^{N} g_i x_{i-s}^T}{\sum_{t=1}^{N} g_t}\left(\sum_{t=1}^{N} g_t x_t - \sum_{k=1}^{K}\alpha_k \sum_{t=1}^{N} g_t x_{t-k}\right) + \sum_{k=1}^{K}\alpha_k \sum_{i=1}^{N} g_i x_{i-s}^T x_{i-k} = \sum_{i=1}^{N} g_i x_{i-s}^T x_i$$

$$s = 1,\ldots,K$$

Для упрощения записи введём обозначения:

$$\sum_{\gamma} g_\gamma = r, \quad \sum_{\gamma} g_\gamma x_{\gamma-l} = r_l, \quad \sum_{\gamma} g_\gamma x_{\gamma-l_1}^T x_{\gamma-l_2} = r_{l_1,l_2},$$

и тогда система примет вид:

$$\frac{r_s}{r}\left(r_0 - \sum_{k=1}^{K}\alpha_k r_k\right) + \sum_{k=1}^{K}\alpha_k r_{s,k} = r_{s,0} \quad s = 1,\ldots,K$$

Домножая на r и перенося свободные слагаемые в правую часть, в итоге получаем

$$\sum_{k=1}^{K}\alpha_k \left(r_{s,k} r - r_s r_k\right) = r_{s,0} r - r_s r_0 \quad s = 1,\ldots,K \qquad (8)$$

То есть искомые коэффициенты авторегрессии являются решением системы линейных алгебраических уравнений (8) с симметрической матрицей.

Если известны коэффициенты регрессии (а следовательно, и ошибка), то можно найти оптимальные параметры так же, как это сделано в [3].

Итак, выпишем все формулы, дающие оценку «новым» параметрам через «старые»[2].

[2] «Старые» параметры входят в указанное выражение неявно, будучи используемыми при подсчёте значений вероятностей $\{g_{ij}\}$ через формулу Байеса.

$$\beta^{new} = \frac{\sum_{i=1}^{N} g_i}{N},$$

$$A\alpha^{new} = b,$$

$$\eta^{new} = \frac{\sum_{i=1}^{N} g_i (x_i - \sum_{k=1}^{K} \alpha_k^{new} x_{i-k})}{\sum_{t=1}^{N} g_i}$$

$$\Lambda^{new} = \frac{\sum_{i=1}^{N} g_i (x_i - \sum_{k=1}^{K} \alpha_k^{new} x_{i-k} - \eta^{new})(x_i - \sum_{k=1}^{K} \alpha_k^{new} x_{i-k} -$$

$$\sum_{t=1}^{N} g_i$$

Изложенный метод позволяет найти параметры, при которых достигается локальный максимум функции правдоподобия. Сходимость и точность решения во многом зависят от выбора начального приближения параметров. Конечно, можно выбрать начальные значения случайным образом, но на практике такой подход приводит к тому, что различные эксперименты (включающие в себя обучение) приводят к различным значениям точности системы. При использовании EM-алгоритма более пригодным оказывается следующий подход: сначала в выборке находятся k классов таких, что расстояние между объектами класса до его центра меньше, чем расстояние до центров всех прочих классов. Потом за начальное приближение математического ожидания берётся среднее значение каждого класса, а за дисперсию (в предположении о некоррелируемости случайных величин) — соответствующее ему среднеквадратичное отклонение. Построение соответствующих классов можно проводить на основе алгоритма k-средних (k-Means).

Для смеси нормальных распределений авторегрессии описанный выше метод организации начального приближения использовать неприемлемо — данные линейно связаны друг с другом. Однако если использовать вместо этих данных ошибку авторегрессии из представления (2), то подход, описанный выше, станет корректным. Остаётся вопрос — как найти начальные коэффициенты регрессии? Здесь можно предложить следующую эвристику: для начала также находим k классов в исходной выборке. Далее для объектов каждого класса независимо от других классов производим «нормализацию»: вычитаем их среднее значение. Предположим, что получена выборка временно̀го ряда, сформированная нормальным процессом авторегрессии с нулевым средним. Тогда коэффициенты этой авторегрессии будут искомыми коэффициентами. Для нахождения коэффициентов используем уравнения Юла-Уолкера (Yule-Walker) [4] , возникающие как следствие минимизации ошибки линейного предсказания.

Критерий остановки описанной выше итерационной процедуры может быть двух типов:

1. Близость по какой-либо метрике оптимизируемых параметров, например,

   при заданной точности δ критерием может являться $\left\| \eta_{new}^{j} - \eta_{old}^{j} \right\|^2 < \delta^2$

2. Стабилизация функции правдоподобия.

Как показали эксперименты, в отличие от первого критерия второй позволяет добиться требуемой точности решения за существенно меньшее число итераций.

Следует также отметить, что число компонент в смеси является структурным параметром [6], поэтому какие-то компоненты могут плохо описывать реальное распределение. В этом случае их априорная вероятность оказывается малой, и их необходимо удалить. Задача оценки оптимального числа компонент смеси для произвольной выборки некорректна.

Решающее правило

В результате обучения формируется s языковых моделей $\{P_1, P_2, \ldots, P_s\}$, каждая из которых характеризуется своим набором компонент смеси: $P_i \sim (p_1^i, p_2^i, \ldots, p_{M_i}^i)$. Язык, которому соответствует i-ая модель, считается распознанным, если правдоподобие i-ой модели максимально.

Логарифм функции правдоподобия имеет вид:

$$MLE_i = \log \prod_{k=1}^{N} P_i(x_k) = \sum_{k=1}^{N} \log P_i(x_k) = \sum_{k=1}^{N} \log \sum_{j=1}^{M_i} \alpha_j^i p_j^i(x_k) \quad i = 1, 2, \ldots, s$$

В этих обозначения решающее правило формально выглядит следующим:

$$i^* = \arg\max_{i=1,2,\ldots,s} MLE_i$$

Результаты

Ниже в таблице приведена точность распознавания пяти языков для трёх описанных выше языковых моделей. Каждая модель включает в себя 50 кластеров.

Все обучающие базы состоят из звуковых файлов с частотой оцифровки 8 КГц и разрядностью 16 бит; общий объём каждой базы — 500 Мб. Данные получены в различных каналах записи: микрофон, цифровая телефонная линия с μ-law кодированием, аналоговый телефон.

Название языка	MFCC	SDC	MFCC + AR
Арабский	95%	78%	91%
Английский	66%	23%	70%
Китайский	92%	54%	91%
Русский	80%	49%	83%
Турецкий	51%	73%	43%
Общая точность:	**76,8%**	**55,4%**	**75,6%**

Эксперименты показали, что авторегрессионная модель языка является менее чувствительной по отношению к каналу записи при незначительном падении точности распознавания, по сравнению с MFCC моделью. Несоответствия искажений, вносимых каналами, компенсируются динамическими свойствами акустических признаков речевого сигнала.

Точность идентификации на основе SDC модели существенно меньше точности, указанной в [5].

Заключение

В данной работе мы рассмотрели и дали сравнительный анализ трём типам акустических моделей речи, используемых для задачи автоматической идентификации языка: кепстральные мел-коэффициенты (MFCC), смещённый дельта кепстр (SDC) и авторегрессионная модель кепстральных коэффициентов (MFCC+AR). Параметры смеси гауссовских распределений, описывающей языковую модель, настраивались посредством обучения на базе размером 500 Мб. Результаты распознавания получены для моделей, состоящих из 50 кластеров. Увеличение числа кластеров ведёт к повышению точности, однако временные затраты на обучение при этом резко возрастают.

Эксперименты показали, что качество системы во многом зависит от однородности искажений всех полученных данных и ухудшается, если данные получены из различных каналов связи. Учёт динамических свойств речевых признаков позволяет повысить устойчивость к каналу при распознавании языка.

Литература

1. *Аграновский А.В., Зулкарнеев М.Ю., Леднов Д.А., Можаев О.Г.* Автоматическая идентификация языка// Искусственный интеллект, № 4, 2002, изд. НАН Украины, Донецк, 2002. С. 142–150.

2. *T. Schultz, Q. Jin, K. Laskowski, A. Tribble, and A. Waibel,* “Speaker, Accent, and Language Identification Using Multilingual Phone Strings”, HLT 2002, San Diego, California, March 2002.

3. *Jeff A. Bilmes* «A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models», International Computer Science Institute Berkeley CA, April 1998.

4. *Gidon Eshel* «The Yule Walker Equations for the AR Coefficients».

5. *Pedro A.* Torres-Carrasquillo «Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features», Proc. Int’l Conf. Spoken Language Processing, Denver, Sep. 2002.

6. *Моттль В.В., Мучник И.Б.* Скрытые Марковские модели в структурном анализе сигналов. М.: ФИЗМАТЛИТ, 1999.

Леднов Дмитрий Анатольевич —

кандидат технических наук, старший научный сотрудник, руководитель отдела речевых технологий ООО «Стел — Компьютерные Системы», окончил Казахский государственный университет, г. Алма-Ата, защитил кандидатскую диссертацию в Ростовском государственном университете, г.Ростов-на-Дону.

Иващенко Юрий Станиславович —

студент 6-го курса Московского института радиотехники, электроники и автоматики, сотрудник отдела речевых технологий ООО «Стел — Компьютерные Системы».

Любимов Николай Андреевич —

студент 5-го курса факультета вычислительной математики и кибернетики МГУ, сотрудник отдела речевых технологий ООО «Стел — Компьютерные Системы».