УДК 004.934.1’1

А.С. Алексеев, Е.Е. Федоров
Институт проблем искусственного интеллекта, г. Донецк, Украина,
fee@iai.donetsk.ua, gotletter@gmail.com

Количественный анализ систем признаков и методов идентификации

Для создания системы идентификации диктора был проведен количественный анализ системы признаков, основанной на линейном предсказании, и системы признаков, основанной на нормированном количестве импульсов равной длины, которые используются в методе, базирующемся на мере различимости Атала, и алгоритме DTW.

Постановка проблемы. В настоящее время актуальной является разработка систем, предназначенных для идентификации диктора. Эти системы имеют широкую область применения – криминалистика (фоноскопическая экспертиза), криптография, охранные системы и др. При разработке таких систем важную роль играет выбор системы признаков и методов идентификации, использующих эти признаки.

Нерешенные ранее проблемы. В работах [1-3] приведены системы идентификации, дающие в большинстве случаев вероятность распознавания ниже 90 %.

Цель и задачи исследования. Для построения системы идентификации необходимо определить систему признаков и использующий ее метод путем количественного анализа.

Решение задачи. В статье рассматриваются:

- системы признаков, основанные на линейном предсказании;
- система признаков, которая основана на нормированном количестве импульсов равной длины;
- метод идентификации, основанный на мере различимости Атала;
- метод идентификации, основанный на алгоритме динамического искажения времени DTW.

Для этих признаков и методов проведен количественный анализ.

Традиционно в качестве системы признаков выбираются признаки, вычисленные с помощью метода кодирования с линейным предсказанием [1-3].

Передаточная функция линейной системы, моделирующей речевой тракт человека, представлена в виде

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \qquad (1)$$

где $S(z)$ – сигнал на выходе линейной системы; $U(z)$ – сигнал возбуждения, поступающий на вход системы, p – порядок линейного предсказателя (обратного фильтра), коэффициент усиления G и коэффициенты цифрового фильтра $\{a_k\}$.

Эта система возбуждается импульсной последовательностью для вокализованных звуков речи и шумом для невокализованных.

Для системы (1) отсчет речевого сигнала $s(n)$ связан с сигналом возбуждения $u(n)$ простым разностным уравнением

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n). \qquad (2)$$

Линейный предсказатель с коэффициентами a_k определяется как система, на выходе которой будет

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k). \quad (3)$$

Передаточная функция фильтра-предсказателя p-го порядка представляет собой полином вида

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k}. \quad (4)$$

В качестве первой системы признаков выступают коэффициенты линейного предсказания (коэффициенты фильтра) α_k, вычисляемые по автокорреляционному методу согласно алгоритму Дарбина [1].

Пусть $R_n(i)$ – автокорреляционная функция; n – номер сегмента речевого сигнала; i – порядок линейного предсказателя; $\alpha_j^{(i)}$ – j-й коэффициент линейного предсказателя порядка i; k_i – i-й коэффициент отражения; $E^{(i)}$ – среднеквадратичная погрешность предсказания для линейного предсказателя порядка i; p – порядок предсказателя.

Коэффициенты линейного предсказания $\alpha_j^{(i)}$, согласно алгоритму Дарбина, вычисляются следующим образом:

$$E_n^{(0)} := R_n(0), \quad (5)$$

$$k_i := \left[R_n(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_n(i-j) \right] \Big/ E_n^{(i-1)}, 1 \le i \le p, \quad (6)$$

$$\alpha_i^{(i)} := k_i, \quad (7)$$

$$\alpha_j^{(i)} := \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}, 1 \le j \le i-1, \quad (8)$$

$$E_n^{(i)} := (1 - k_i^2) E_n^{(i-1)}. \quad (9)$$

Окончательное решение принимает вид

$$\alpha_j := \alpha_j^{(p)}, 1 \le j \le p. \quad (10)$$

Для устойчивости линейной системы (1) требуется выполнение условия

$$-1 \le k_i \le 1. \quad (11)$$

$E_1 = (\alpha_1, ..., \alpha_p)$ – вектор признаков, основанный на коэффициентах линейного предсказания.

К системам признаков, основанным на линейном предсказании, относятся также кепстральные коэффициенты и площади акустической трубы [1].

Кепстральные коэффициенты $\hat{h}(n)$ вычисляются через коэффициенты линейного предсказания:

$$\hat{h}(n) = \alpha_n + \sum_{k=1}^{n-1} \frac{k}{n} \cdot \hat{h}(k) \alpha_{n-k}, n \in \overline{1, p}. \quad (12)$$

$E_2 = (\hat{h}(1),...,\hat{h}(p))$ – вектор признаков, основанный на кепстральных коэффициентах.

Функция площадей акустической трубы A_i вычисляется через коэффициенты отражения:

$$A_{i+1} = \frac{1-k_i}{1+k_i} A_i,\ A_1 = 1, i \in \overline{1, p}. \tag{13}$$

$E_3 = (A_2,..., A_{p+1})$ – вектор признаков, основанный на площадях акустической трубы.

В отличие от традиционных систем признаков E_1, E_3, E_3, авторы статьи предлагают в качестве признаков использовать *нормированное количество импульсов равной длины*, использовавшееся ранее только в системах распознавания [4].

Пусть $X = \{x_1,..., x_n\}$ – оцифрованный звуковой сигнал.

Сигнал подвергается *M*-кратному сглаживанию фильтром.

$$y_j = \frac{x_{j-1} + x_j + x_{j+1}}{3}. \tag{14}$$

Далее вычисляется разность исходного и сглаженного сигналов

$$y_j^{(0)} = x_j - y_j. \tag{15}$$

Затем определяется длина импульса *z*:

$$z = \begin{cases} l, 2 \le l < 20 \\ 20 + \left[\frac{l-20}{6}\right], 20 \le l < 50 \\ 25 + \left[\frac{l-50}{10}\right], 50 \le l < 90 \\ 29, l \ge 90 \end{cases},\ z \in [2,29], \tag{16}$$

где *l* – число отсчетов между двумя соседними локальными максимумами сигнала

$$(y_{j-1}^{(0)} < y_j^{(0)} \ge y_{j+1}^{(0)})$$

и количество импульсов $n_k^{(0)}$ длины *k*+1.

После чего в цикле $i \in \overline{1,N}$ сигнал подвергается двукратному сглаживанию фильтром (17):

$$y_j^{(i)} = \frac{y_{j-1}^{(i-1)} + y_j^{(i-1)} + y_{j+1}^{(i-1)}}{3} \tag{17}$$

и рассчитывается количество импульсов $n_k^{(i)}$.

По завершении цикла вычисляется общее количество импульсов одинаковой длины

$$n_k = \sum_{i=0}^{N} n_k^{(i)}, k \in \overline{1,28} \tag{18}$$

и количество всех импульсов

$$n = \sum_{k=1}^{28} n_k. \tag{19}$$

$E_4 = (e_1,...,e_{28})$ – вектор признаков, где $e_k = n_k / n$ – нормированное количество импульсов длины $k + 1$.

Для идентификации диктора обычно используют сложные и устойчивые к различным аномалиям меры различимости. Традиционно применяется *мера различимости, предложенная Аталом* [1-3]. Эта мера может быть получена следующим образом. Пусть x представляет собой вектор-столбец измеренных значений входного сигнала размерностью L, причем элементом x является k-е измеренное значение. Предполагается, что совместная функция плотности вероятности измеренных значений для i-го диктора представляет собой многомерное распределение Гаусса со средним значением m_i и ковариационной матрицей W_i. Таким образом, L-мерная плотность распределения Гаусса для x имеет вид

$$g_i(x) = (2\pi)^{-\frac{1}{2}} \cdot |W_t|^{-\frac{1}{2}} \cdot \exp[-\frac{1}{2} \cdot (x - m_i)^t \cdot W_i^{-1} \cdot (x - m_i)], \quad (20)$$

где W_i^{-1} – матрица, обратная W_i; $| W_i |$ – детерминант W_i; t – транспонирование вектора.

Решающее правило, минимизирующее вероятность ошибки, состоит в том, что вектор измеренных значений x следует отнести к классу i, если

$$p_i \cdot g_i(x) \geq p_j \cdot g_j(x), i \neq j, \quad (21)$$

где p_i – априорная вероятность принадлежности вектора x к классу i.

Поскольку $\ln y$ – монотонно возрастающая функция своего аргумента, решающее правило (21) можно значительно упростить, переписав в виде

$$d_i(x) = \frac{1}{2} \cdot (x - m_i)^t \cdot W_i^{-1} \cdot (x - m_i) + \frac{1}{2} \ln|W_i| = \ln p_i \leq d_j(x), i \neq j. \quad (22)$$

Последние два члена в правой части (22) не зависят от вектора x, и поэтому можно считать, что они представляют собой смещение i-го класса. Для большинства практически важных случаев установлено, что решающее правило со смещением в правой части не имеет преимуществ перед решающим правилом, основанным только на первом члене (22). Таким образом, функцию различимости можно определить как

$$d_i(x) = (x - m_i)^t \cdot W_i^{-1} \cdot (x - m_i). \quad (23)$$

Решающее правило предполагает вычисление вектора средних и ковариационной матрицы для каждого класса i на множестве решения. Вектор средних и ковариационная матрица определяются по обучающей последовательности $x_i(n)$ векторов, принадлежащих i-му классу:

$$m_i = \frac{1}{N_i} \cdot \sum_{n=1}^{N_i} x_i(n), \quad (24)$$

$$W_i = \frac{1}{N_i} \cdot \sum_{n=1}^{N_i} x_i(n) \cdot x_i^t(n) - m_i \cdot m_i^t. \quad (25)$$

Для эффективного поиска минимального расхождения между входным сигналом X и эталоном E может также использоваться *алгоритм динамического искажения времени DTW* [4], [5]. Его ключевая идея заключается в том, что в точке (i, j) просто продолжается самый близкий маршрут сравнения из $(i - 1, j - 1)$, $(i - 1, j)$ или $(i, j - 1)$.

Предварительно формируется темпоральная транскрипция обучаемого или распознаваемого слова, т.е., используя меры сходства, производится сопоставление вектора признаков каждого сегмента слова с эталонами звуков и замена вектора признака сегмента на индекс ближайшего ему эталона. Это позволяет компактно представить слово.

Пусть C_{ij} – глобальное расхождение от точки (i, j), D_{ij} – локальное. Тогда

$$C_{ij} = D_{ij} + \min(C_{i-1,j}, C_{i,j-1}, C_{i-1,j-1}), \tag{26}$$

$$D_{ij} = \sqrt{\sum_s (x_{is} - e_{js})^2}, \tag{27}$$

где x_{is} – s-й признак i-го сегмента распознаваемого слова, e_{ji} – s-й признак j-го сегмента эталона звука.

Начальное условие: $C_{11} = D_{11}$.

Результатом работы алгоритма DTW является получение C_{NN}.

Количественный анализ систем идентификации был проведен авторами статьи следующим образом. Каждый диктор произносил 5 раз слово «Саша». Первые четыре реализации рассматривались как эталон, пятая реализация сравнивалась с эталонами всех дикторов.

По результатам идентификации для каждой распознаваемой реализации выбирались 5 ближайших к ней эталонов. Идентификация считалась успешной, если среди этих пяти встречались эталоны, соответствующие именно тому диктору, которому в действительности соответствовала распознаваемая реализация (удовлетворительный результат). Наибольший интерес представляют случаи, когда ближайшим являлся эталон, соответствующий требуемому диктору, – это означало, что диктор однозначно определен (отличный результат). Результаты исследований приведены в табл. 1.

Таблица 1 – Результаты численного сравнения методов идентификации диктора

Система признаков	Метод идентификации	К-во эталонов	Успешное распознавание	Однозначная идентификация
Нормированное количество импульсов равной длины	DTW	60	92,31 %	57,69 %
Коэффициенты линейного предсказания	мера Атала	60	13,79 %	3,45 %
Кепстральные коэффициенты	мера Атала	60	12,07 %	3,45 %
Функция площади	мера Атала	60	50,00 %	15,52 %
Функция площади	DTW	60	65,52 %	37,93 %

Согласно табл. 1, наилучшие результаты дает сочетание нормированного количества импульсов равной длины с алгоритмом DTW – вероятность идентификации составляет 92 %. Аналогичные зарубежные системы дают вероятность примерно 70 – 95 %.

Выводы

Новизна. В статье рассматривались системы признаков, как традиционные (основаны на линейном предсказании), так и предложенные авторами (основаны на нормированном количестве импульсов равной длины), которые используются при идентификации диктора. Эти признаки использовались в методе, базирующимся на мере различимости Атала и алгоритме DTW. Для этих признаков и методов проведен количественный анализ, в результате которого для идентификации диктора было выбрано сочетание нормированного количества импульсов равной длины с алгоритмом DTW.

Практическое значение. Основные положения работы были использованы при разработке системы идентификации диктора, которая может использоваться в криминалистике (фоноскопическая экспертиза) и охранных системах.

Литература

1. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. – М.: Радио и связь, 1981. – 496 с.
2. Атал Б.С. Автоматическое опознавание дикторов по голосам // ТИИЭР. – 1976. – Т. 64, № 4. – С. 48-66.
3. Galoonov V.I., Gramnitski S.N., Romashov N.A. VQ and GMM combination for text-independent speaker recognition on telephone channel // SPECOM'2002. – P. 57-60.
4. Дорохин О.А., Засыпкин А.В., Червин Н.А., Шелепов В.Ю. О некоторых подходах к проблеме компьютерного распознавания устной речи // Труды Междунар. конф. «Знание-Диалог-Решение» (KDS 97). – Т. 1. – Ялта. – 1997. – С. 234-240.
5. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. – К.: Наук. думка, 1987. – 261 с.

О.С. Алексєєв, Є.Є. Федоров

Кількісний аналіз систем ознак і методів ідентифікації

Для створення системи ідентифікації диктора в статті був проведений кількісний аналіз системи ознак, заснованої на лінійному завбаченні, і системи ознак, заснованої на нормованій кількості імпульсів рівної довжини, що використовуються в методі, заснованому на мірі розрізнення Атала, і алгоритмі DTW.

A. Alexeev, E. Fedorov

Quantification of Systems of Features and Methods of Identification

For creation of a system of identification of the speaker in the article the quantification of a system of features, founded on a linear prediction was conducted, and the systems of features, founded on normalized quantity of pulses of equal length, which one will be used in a method, founded on a measure of a conspicuity Atal and algorithm DTW.

Статья поступила в редакцию 04.07.2005.