АЛГОРИТМ ИСПОЛЬЗОВАНИЯ ГАУССОВЫХ СМЕСЕЙ ДЛЯ ИДЕНТИФИКАЦИИ ДИКТОРА ПО ГОЛОСУ В ТЕХНИЧЕСКИХ СИСТЕМАХ

Д. В. Пекарь, С. Г. Тихоненко

Белорусский государственный университет, Минск, Беларусь

Одним из способов решения задачи идентификации человека является анализ его биометрических параметров, своего рода биологического паспорта человека, так как уже природой заложена его уникальность.

ИЗВЛЕЧЕНИЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ РЕЧЕВОГО СИГНАЛА

Мелкепстральный анализ, далее MFCC, является эффективным способом извлечения информации об особенностях речеобразования [1] из звукового сигнала и сжатого его представления. Известно [2], что на интервале 8-20 мс форма речевого сигнала претерпевает незначительные изменения, поэтому его характеристики можно считать постоянными в пределах временного окна. При определении мелкепстральных коэффициентов на первом этапе вычисляется спектр:

$$x(k) = \sum_{n=0}^{N-1} w(n)\, x(n)\, \exp(-j\, 2nkn / N), \quad (1)$$

где x(n) - отсчеты сигнала, N- размер окна, w(n)- весовое окно Хэмминга. Далее вычисляется амплитудный спектр, который затем пропускается через банк треугольных фильтров с центральными частотами f_c(m), где $m = 1,\ldots,k$- номер фильтра, K - количество фильтров. Банк фильтров описывается выражением (2):

$$H(k, m) = \begin{cases} 0, & f(k) < f_c(m-1) \\ \dfrac{2(f(k) - fc(m-1))}{(fc(m+1) - fc(m-1))(f(m) - fc(m-1))}, & fc(m-1) < f(k) < fc(m) \\ \dfrac{2(fc(m+1) - f(k))}{(fc(m+1) - fc(m-1))(f(m+1) - fc(m))}, & fc(m) < f(k) < fc(m+1) \\ 0, & f(k) > fc(m+1) \end{cases}$$

где $f(k)$ - k-ая компонента спектра. Затем логарифмируется отфильтрованный спектр:

$$X(m) = \ln\left(\sum_{k=0}^{N-1} |x(k)|H(k, m)\right) \quad (10)$$

Далее подвергается дискретному косинусному преобразованию X(m):

$$c(l) = \sum_{k=i}^{K} X(k) \cos\left(l \cdot (k - \tfrac{1}{2})\right), \quad (4)$$

где c(l) - l-ый MFCC коэффициент, $l = 1,..., K$.

СМЕСЬ ГАУССОВЫХ РАСПРЕДЕЛЕНИЙ

Модель смеси гауссовых распределений X, далее СГР, является параметрической вероятностной функцией, представляющей собой взвешенную сумму функций гауссовых распределений (10):

$$P(\bar{x} \mid X) = \sum_{i=1}^{M} w_i p_i(\bar{x}), \quad (5)$$

где M - количество компонент $P_i(x)$, определяет порядок модели; w_t - вес i-ой компоненты; веса удовлетворяют условию: $\sum_{i}^{M} w_i = 1$.

Компонента смеси имеет вид:

$$P_i(x) = \frac{1}{(2\pi)^{D/2} \sqrt{|\Sigma_i|}} \exp\left[-\frac{1}{2}(x - \mu)^T \Sigma_i^{-1} (x)\right], \quad (6)$$

где Σ_i - ковариационная матрица i-ой компоненты, μ - вектор средних i-ой компоненты, D - размерность вектора X; $\{w, \Sigma, \mu\}$ - параметры, определяющие модель X. СГР используется как модель функции распределения акустических характеристик в речевых системах [3].

Идентификация диктора по образцу речевого сигнала эквивалентна нахождению модели, максимизирующей апостериорную вероятность:

$$X_{ML} = \arg\max_{S} p(X \mid X), \quad (7)$$

где $1 < s < S$ - число дикторов. При применении байесовского правила к (7):

$$X_{ML} = \arg\max_{S} \frac{P(x \mid X) P(X_S)}{p(X)}, \quad (8)$$

где $P(X_S)$ - априорная вероятность появления определенной модели диктора, получим $P(X_1) = P(X_2) = ... = P(X_S)$, $P(X)$, т.е. априорная вероятность появления определенного вектора признаков принимается

равной для всех моделей [4]. Тогда (8) сводится к максимизации функции правдоподобия:

$$Д = \underset{S}{\operatorname{argmax}} P(X \setminus Д). \quad (9)$$

АДАПТАЦИЯ МОДЕЛИ СМЕСИ ГАУССОВЫХ РАСПРЕДЕЛЕНИЙ

Адаптация модели - нахождение таких параметром модели, при которых она бы наилучшим образом, в некотором смысле, описывала распределение значений вектора признаков для заданного диктора. Начиная с начального приближения параметров модели, находят новые, более адаптированные, которые служат начальным приближением для следующей итерации. Алгоритм выполняется пока истинно (10) или не превышено максимально допустимое число итераций:

$$P(X\setminus Д) > P(X \setminus Д). \quad (10)$$

На каждой итерации для i-ой компоненты осуществляются операции (11), (12), (13):

$$wn^{ew} = \frac{1}{T}\sum_{t=1}^{t} P(i\setminus xt\ Д), \quad (11)$$

$$U^{new} = \frac{\sum_{t=1} P(i \setminus x_t, Д)\, x_t}{\sum_{t=1} P(i \setminus x_t, Д)}, \quad (12)$$

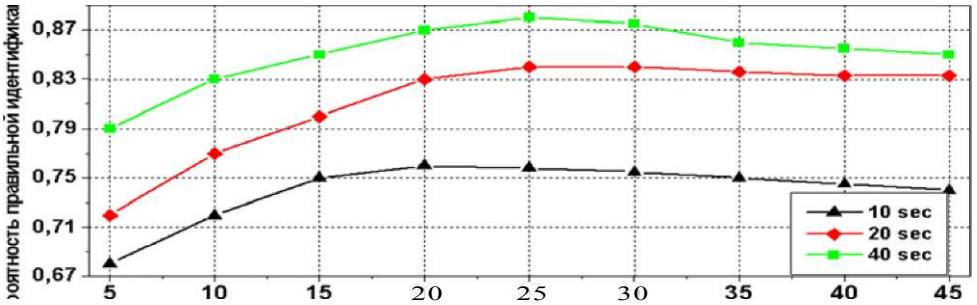
$$Z_i = \frac{\sum_{t=1}^{?} P(i \setminus x, Д)(x - UJrx\ x - jf))}{\sum_{t=1}^{?} P(i \setminus Xt, Д)}, \quad (13)$$

где $1 < t <)$, $)$- число точек, $P(i \setminus x_f Д) = \frac{\text{тм}^{lPl}(x \wedge}{\sum_{m=1} w_m p_m(x_f)}$. (14)

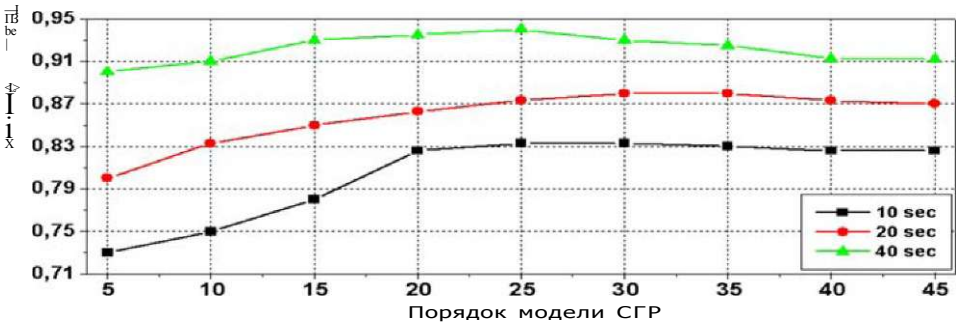
Для начального приближения параметров модели использовался алгоритм кластеризации ^-средних.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

В эксперименте (рис. 1, рис. 2, рис. 3) приняло участие 5 женщин и 5 мужчин, которые произносили по 50 различных слов для обучения и по 30, отличных от первых, слов для эксперимента.

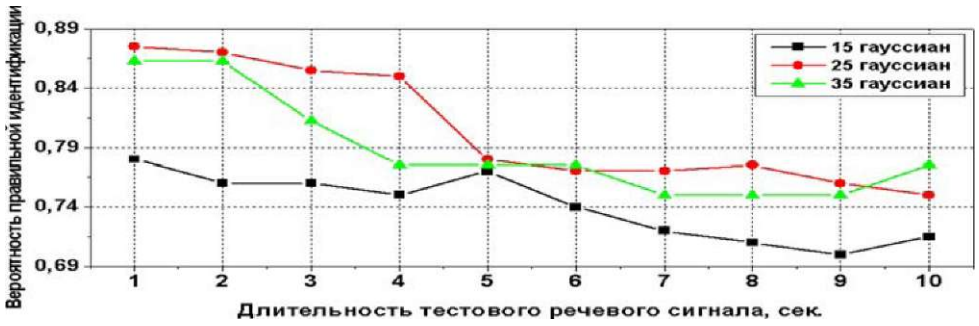


а

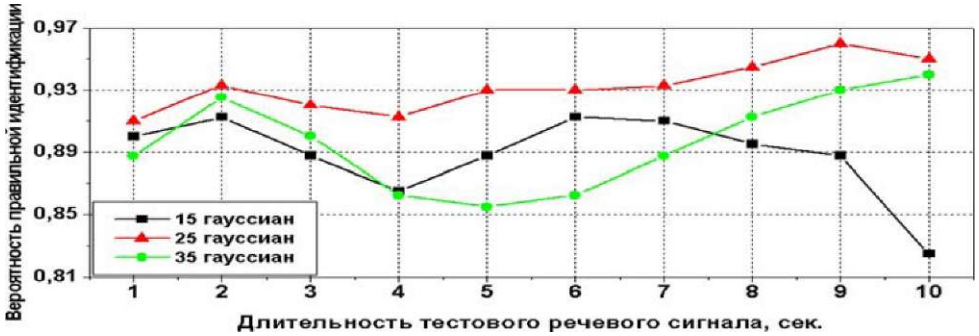


б

Рис. 1. Зависимость достоверности идентификации от порядка модели:
а - 15 MFCC коэффициентов, *б* - 25 MFCC коэффициентов.
Увеличение длительности обучающего фрагмента позволяет накопить больше статистических данных о дикторе и повысить достоверность идентификации



а



б

Рис. 2. Зависимость достоверности идентификации от длины тестовой выборки:
а - 15 MFCC коэффициентов, *б* - 25 MFCC коэффициентов

Применялись речевые фрагменты с длинами 10, 20 и 40 секунд (рис. 1) для обучения, модели с порядками от 5 до 45 с шагом в 5 компонент, модели с полной и диагональной ковариационной матрицей (рис. 3), тестовые речевые фрагменты длиной от 1 до 10 секунд (рис. 2), длина вектора признаков составляла 15 и 25 MFCC-коэффициентов.

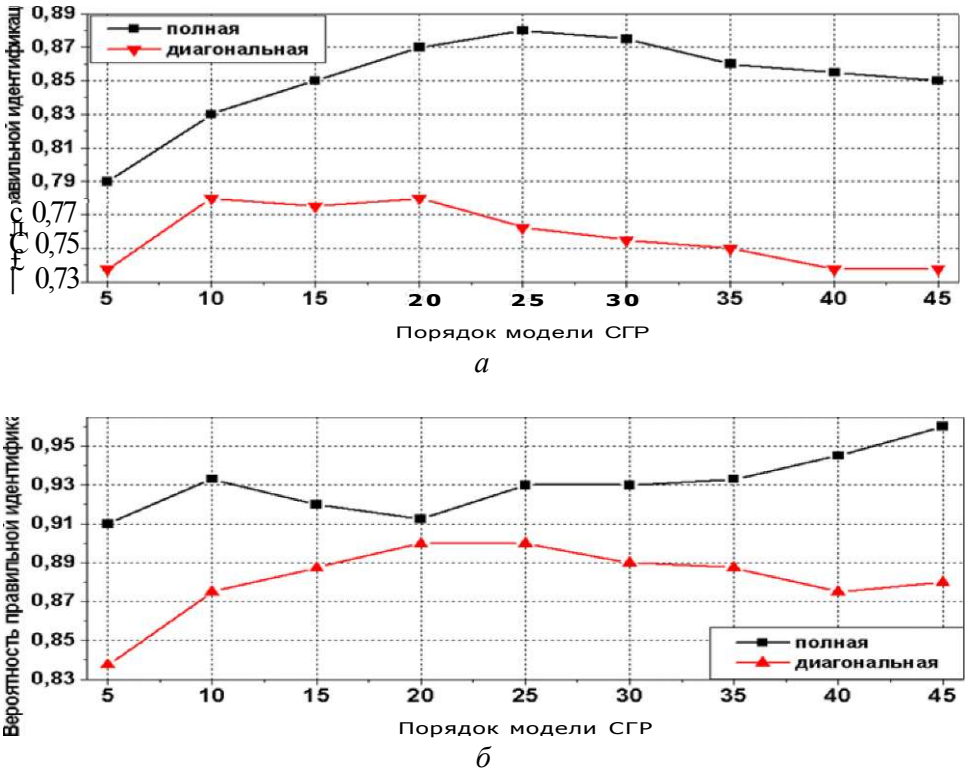


Рис. 3. Зависимость достоверности идентификации для различных типов ковариационных матриц:

а - 15 MFCC коэффициентов, *б* - 25 MFCC коэффициентов.

Полная ковариационная матрица дает лучшие результаты, так как хранит в себе связи между элементами, в тоже время более требовательна к вычислительным ресурсам

Литература

1. *Honda Masaaki,* Human Speech Production Mechanisms / Masaaki Honda. 2003.
2. Интернет-адрес: http://ieeexplore.ieee.org/iel5/10632/33566].
3. Интернет-адрес: http://www.ll.mit.edu/mission/communications/ist/publications].
4. *Bryan, L.* An efficient scoring algorithm for Gaussian Mixture Model based speaker identification // Center for Robust Speech Systems, Department of Electrical Engineering, the University of Texas, Dallas, USA. 1998.