УДК 004.891.3

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ЧЕЛОВЕКА В ДИАЛОГОВЫХ СИСТЕМАХ

Ронкер В.Ю.,
научный руководитель д-р техн. наук, проф. Медведев А.В.
Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева

Биометрические системы распознавания становятся всё более популярны в современном мире. Они являются наиболее "естественным" средством идентификации людей. Вместо того чтобы помнить пароли и PIN-коды (которые могут быть украдены или забыты) или использовать ручные подписи (которые можно подделать), биометрические данные, такие как отпечатки пальцев, голос и лицо являются индивидуальными особенностями человека и, следовательно, не могут быть легко украдены или копированы, а так же не могут быть забыты [1]. Самые простые для распознавания биометрические данные, которые наиболее часто используются и широко распространены в обществе - человеческая речь. Таким образом, решение задачи идентификации для систем распознавания применимо в тех технологиях, которые используют человеческую речь, чтобы распознать, определить или проверить человека [2].

Работа систем распознавания содержит два основных этапа: регистрация пользователей в системе и сам процесс распознавания (попытка идентификации или верификации). Пользователи предварительно регистрируются в системе, записав свои голоса. Образец голоса каждого диктора обрабатывается с целью извлечения признаков, которые могут быть использованы для распознавания. На основе извлечённых признаков строятся модели пользователей. Модель представляет собой некоторую структуру, позволяющую при данных признаках оценить степень подобия либо сразу принять решение.

В случае верификации пользователь пытается войти в систему, предъявляя идентификатор и образец голоса. Признаки, извлечённые из предъявленного образца, сравниваются с соответствующей моделью, сохранённой в базе, а так-же, возможно, с референтной моделью, представляющей фиксированное множество некоторых пользователей, либо наиболее близких к данному голосу. Результат сравнивается с заданным порогом и выдаётся положительное или отрицательное решение о допуске.

Во время процесса идентификации также происходит извлечение признаков из предъявленного образца, которые затем сравниваются с моделями всех зарегистрированных в системе пользователей либо предварительно отобранных.

Для выделения признаков выполняются алгоритмы, использующие мэл-частотные кепстральные коэффициенты [3]. Данный подход является одним из самых распространённых как в системах распознавания дикторов, так и в системах распознавания речи.

На вход алгоритма подаётся последовательность отсчётов участка сигнала, исследуемого на данной итерации, $x_0, \dots, x_{N-1}$. К данной последовательности применяется весовая функция и затем дискретное преобразование Фурье. Весовая функция используется для уменьшения искажений в Фурье анализе, вызванных конечностью выборки. На практике в качестве весовой функции часто используется окно Хэммига, которое имеет следующий вид:

$$w_n = 0{,}54 - 0{,}46 \cdot \cos\left(2\pi \frac{n}{N-1}\right), n = 0, \dots, N-1, \quad (1)$$

где N – длина окна, выраженная в отсчётах.

Теперь дискретное преобразование Фурье взвешенного сигнала можно записать в виде

$$X_k = \sum_{n=0}^{N-1} x_n w_n e^{\frac{-2\pi i}{N} kn},\ k = 0, \dots, N-1. \quad (2)$$

Значение индексов k соответствует частотам

$$f_k = \frac{F_s}{N} k, k = 0, \dots, N/2, \quad (3)$$

где F_s — частота дискретизации сигнала.

Полученное представление сигнала в частотной области разбивают на диапазоны с помощью банка (гребёнки) треугольных фильтров. Границы фильтров рассчитывают в шкале *мэл*. Данная шкала является результатом исследований по способности человеческого уха к восприятию звуков на различных частотах.

Перевод в мэл-частотную область осуществляют по формуле

$$B(f) = 1127 \cdot \ln\left(1 + \frac{f}{700}\right).$$

Обратное преобразование выражается как

$$B^{-1}(b) = 700\left(e^{\frac{b}{1127}} - 1\right).$$

Пусть N_{FB} — количество фильтров (обычно используют порядка 24 фильтров), (f_{low}, f_{high}) — исследуемый диапазон частот. Тогда данный диапазон переводят в шкалу мэл, разбивают на N_{FB} равномерно распределённых перекрывающихся диапазона и вычисляют соответствующие границы в области линейных частот. Обозначим через Hm, k — весовые коэффициенты полученных фильтров. Фильтры применяются к квадратам модулей коэффициентов преобразования Фурье. Полученные значения логарифмируются

$$e_m = \ln\left(\sum_{k=0}^{N} |X_k|^2 H_{m,k}\right), m = 0, \dots, N_{FB} - 1.$$

Заключительным этапом в вычислении MFCC коэффициентов является дискретное косинусное преобразование

$$c_i = \sum_{m=0}^{N_{FB}-1} e_m \cos\left(\frac{\pi i (m + 0{,}5)}{N_{FB}}\right), i = 1, \dots, N_{MFCC}.$$

Далее строится модель звукового сигнала. Она представляет собой взвешенную сумму Гауссиан:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x), \quad (4)$$

где λ — модель звукового сигнала, M — количество компонентов модели, w_i — веса компонентов такие, что $\sum_{i=1}^{M} w_i = 1$.

Функция плотности вероятности каждого компонента даётся формулой

$$p_i(x) = \frac{1}{(2\pi)^{D/2} |\sum_i|^{1/2}} exp\left(-\frac{1}{2}(x - \mu_i)^T \sum_i{}^{-1} (x - \mu_i)\right),$$

где D — размерность пространства признаков, μ_i — вектор математического ожидания, $\sum$ — матрица ковариации. Чаще всего в системах, реализующих данную модель, используется диагональная матрица ковариации. Возможно также

использование одной матрицы ковариации для всех компонентов модели диктора или одной матрицы для всех моделей.

Таким образом, для построения модели диктора необходимо определить векторы средних, матрицы ковариации и веса компонентов. Данную задачу решают с помощью EM-алгоритма. На вход подаётся обучающая последовательность векторов $X = \{x_1, \ldots, x_T\}$. Параметры модели инициализируются начальными значениями и затем на каждой итерации алгоритма происходит переоценка параметров.

Для определения начальных параметров обычно используют алгоритм кластеризации такой, как алгоритм К-средних [4]. Построив разбиение множества обучающих векторов на M кластеров, параметры модели могут быть инициализированы следующим образом. Начальные значения μ_i совпадают с центрами кластеров, матрицы ковариации рассчитываются на основе попавших в данный кластер векторов, веса компонентов определяются долей векторов данного кластера среди общего количества обучающих векторов.

Переоценка параметров происходит по следующим формулам:

- вычисление апостериорных вероятностей:

$$p(i|x_t, \lambda) = \frac{w_i p_i(x_t)}{\sum_{k=1}^{M} w_k p_k(x_t)}$$

- вычисление новых параметров модели:

$$w_i = \frac{1}{T}\sum_{t=1}^{T} p(i|x_t, \lambda)\,; \qquad \mu_i = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} p(i|x_t, \lambda)}$$

$$\sum_i = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda)(x_t - \mu_i)(x_t - \mu_i)^T}{\sum_{t=1}^{T} p(i|x_t, \lambda)}$$

Данные шаги повторяются до схождения параметров. В результате мы имеем модели спикеров.

На данный момент разработана программа. В ней реализованы алгоритмы распознавания диктора по данным его речи, которые были описаны выше. В качестве таких данных используются звуковые файлы заданного формата. В качестве результата программа формирует решение задачи распознавания в виде матрицы: строки соответствуют экзаменующим отрывкам речи, столбцы – дикторам, а значения матрицы содержат оценку принадлежности конкретного отрывка конкретному диктору. В дальнейшем предполагается создать пользовательский интерфейс для удобства работы с программой, а также повысить возможность интеграции ее в другие системы.

ЛИТЕРАТУРА

1. A. Jain, A. Ross, and S. Prabhaker, «An introduction to biometric recognition», IEEE Trans. Circuits Systems Video Technol., vol.14, no. 1, pp. 4-20, 2004.
2. D. Reynolds, «An overview of automatic speaker recognition technology», in Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP), 2002, vol. 4, pp. 4072-4075.
3. S. Furui, «Cepstral analysis technique for automatic speaker verification», IEEE Trans. Acoustics Speech Signal Process., vol.29, no. 2, pp. 254-272, 1981.
4. Садыхов Р.Х., Ракуш В.В. Модели гауссовых смесей для верификации диктора по произвольной речи // Доклады БГУИР. Минск, 2003. № 4. С. 95-103.