УДК 004.056+ 004.852

А. С. Заковряшин, П. В. Малинин, А. А. Лепендин

Применение распределений мел-частотных кепстральных коэффициентов для голосовой идентификации личности

A. S. Zakovryashin, P. V. Malinin, A. A. Lependin

Speaker Recognition Using Mel-Frequency Cepstral Coefficient Distributions

Работа посвящена развитию методов распознавания личности на основе голосовых данных. Предложен новый подход к формированию векторов признаков при предварительной обработке голосовых образцов, основанный на построении гистограмм частотных распределений мел-частотных кепстральных коэффициентов. Отличительной особенностью является независимость полученного вектора от длины исходного голосового образца, его относительно малый размер и учет в нем разброса индивидуальных характеристик голосового тракта идентифицируемого субъекта. Разработан программный модуль идентификации личности по голосу на основе предложенного подхода и метода опорных векторов. Программный модуль реализован на языке Matlab с использованием функций пакета Voicebox. Проведено сравнение с традиционно используемыми при решении задачи идентификации дикторов векторами признаков. Тестовые испытания разработанного модуля показали, что предложенный подход к предварительной обработке голосовых данных позволяет достичь относительно низкого значения вероятностей ошибок первого и второго рода и может использоваться при построении эффективных систем речевой идентификации.

Ключевые слова*:* голосовая идентификация личности, вектор признаков, мел-частотные кепстральные коэффициенты, распределение частот.

This paper is devoted to the development of feature extraction methods for speaker recognition. A new approach based on histograms of mel-frequency cepstral coefficient (MFCC) distributions to calculate feature vectors for voice samples is proposed. The resulting vectors appear to be independent of original voice sample length and have relatively small sizes. They incorporate the spread of unique vocal tract related characteristics which can be used as distinctive features for recognition. This approach of voice recognition is implemented in a software module developed for MATLAB environment. A support vector machine method and Voicebox speech processing toolbox for MATLAB are utilized. Results of the developed module test runs are obtained and reported. A comparison of test results with results of traditionally used feature vector based techniques of speaker recognition shows relatively low rates of false acceptance and false match for the proposed approach. Feature vectors based on MFCC distributions can be effectively used in real world voice recognition systems.

Key words: speaker recognition, feature vector, mel-frequency cepstrum coefficients, frequency distribution.

DOI 10.14258/izvasu(2014)1.1-35

Введение. Задача верификации диктора по голосовым данным в настоящее время находит широкое применение при построении безопасных информационных систем. Как правило, ее решение основывается на выявлении индивидуальных акустических характеристик пользователей, которые бы позволили эффективно и точно проводить сравнение образцов голоса, предъявляемых при попытке доступа и сохраняемых в специализированной базе данных.

Как и любой другой биометрический подход, голосовая идентификация не является абсолютно надежной. На ее качество влияют расположение диктора относительно микрофона, состояние его здоровья (наличие или отсутствие хрипа в голосе), характеристики регистрирующего тракта, особенности реализации алгоритмов предварительной обработки сигнала и получения вектора признаков, его характеризующего, применяемый алгоритм идентификации. Таким образом, несмотря на активное развитие систем голосовой идентификации, имеется необходимость в их постепенном совершенствовании.

В настоящей работе предлагается новый подход к формированию вектора признаков, описывающего индивидуальные характеристики голоса диктора. Он основан на применении уже хорошо зарекомендовавшего способа выделения полезной информации об акустиче-

ском сигнале, основанном на вычислении мел-частотных кепстральных коэффициентов (MFCC–Mel Frequiency Cepstral Coefficients) и построении их распределений для фраз произвольной длины. Отличительной особенностью предлагаемого подхода является независимость полученного вектора от длины исходного голосового образца, его относительно малый размер и учет в нем разброса индивидуальных характеристик голосового тракта идентифицируемого субъекта.

1. Получение вектора признаков на основе MFCC. Схема системы идентификации личности на основе голосовых данных реализуется с помощью следующих этапов [1,2]:

1. Уровень обработки сигнала. Выделение признаков, существенных для задачи распознавания и формирование так называемого вектора признаков.

2. Уровень модели. Позволяет путем построения математической модели проводить сопоставление векторов признаков друг с другом и вычислять степени подобия между зарегистрированными признаками и сохраненной моделью.

3. Уровень принятия решений. Проводит принятие конечных решений на основе полученных степеней подобий и, если необходимо, заданных пороговых значений.

К настоящему времени в отрасли сложился типичный алгоритм предварительной обработки акустического сигнала после его записи [3]. Оцифрованный сигнал разбивается на блоки длительностью 25–30 мс (обозначим отсчеты в одном из них $x_0,\ldots, x_{N-1}$). К каждому подобному блоку применяется весовая функция и затем дискретное преобразование Фурье. Примером весовой функции может служить окно Хэмминга:

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

где N — длина окна, выраженная в отсчетах.

Весовая функция используется для уменьшения искажений в Фурье анализе, вызванных конечностью выборки. Тогда дискретное преобразование Фурье взвешенного сигнала можно записать в виде:

$$X_k = \sum_{n=0}^{N-1} x_n w_n \exp\left(-\frac{2\pi i}{N} kn\right). \quad (2)$$

Значения индексов k соответствуют частотам:

$$f_k = \frac{F_s}{N} k, \quad (3)$$

где F_s — частота дискретизации сигнала.

Полученное представление сигнала в частотной области разбивают на диапазоны с помощью банка (гребенки) треугольных фильтров. Границы фильтров рассчитывают в шкале мел. Перевод в мел-частотную область осуществляется по формуле [4]:

$$B(f) = 1127 \cdot \ln\left(1 + \frac{f}{700}\right). \quad (4)$$

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

$$e_m = \ln\left(\sum_{k=0}^{N} |X_k|^2 H_{m,k}\right), \quad m = 0,\ldots,N_{FB}-1. \quad (5)$$

Заключительным этапом в вычислении MFCC коэффициентов является дискретное косинусное преобразование

$$c_i = \sum_{m=0}^{N_{FB}-1} e_m \cos\left(\frac{\pi i(m+0{,}5)}{N_{FB}}\right), \quad i = 1,\ldots,N_{MFCC}. \quad (6)$$

Коэффициент c_0 не используется, так как представляет энергию сигнала. Количество коэффициентов N_{MFCC} на практике выбирают от 12 до 30. На рисунке 1 приведен пример графика мел-кепстральных коэффициентов.

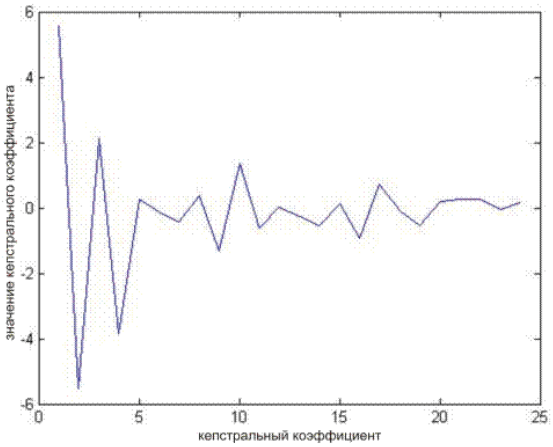

Рис. 1. Пример мел-кепстральных коэффициентов для фразы «один-два-три»

Для типичных акустических сигналов, применяемых при идентификации (коротких парольных фраз типа «один, два, три»), количество блоков разбиения, для которых мы подсчитываем коэффициенты MFCC, является плавающей величиной, зависящей от длительности произнесения фразы. Соответствующие вектора признаков имеют также различную длину и содержат порядка нескольких тысяч кепстральных коэффициентов. В некоторых случаях [1] к этим данным могут добавляться также еще и рассчитанные на основе MFCC значения первых и вторых производ-

ных по времени, что еще больше увеличивает длины векторов.

Существует несколько подходов к фиксации и уменьшению длины результирующего вектора признаков:

- размер окна для разбиения сигнала брать не фиксированной длины для всех образцов, а разбивать их на фиксированное количество окон, длины, рассчитываемой для каждого образца;
- не разбивать сигнал на окна, а получать вектор признаков значений мел-частотных кепстральных коэффициентов для всего сигнала. Для всех образцов длина вектора признаков будет равна заданному количеству кепстральных коэффициентов. Описание сигнала становится крайне грубым;
- приводить все образцы на этапе предобработки к одной длине. Этот метод является неприемлемым в задаче распознавания диктора, так как вносит искажения в исходный сигнал.

В данной работе был предложен новый способ формирования вектора признаков для образца речевого сигнала на основе частотного распределения значений, полученных при применении алгоритма мел-частотных кепстральных коэффициентов. Будем работать с набором векторов кепстральных коэффициентов, размером MxN_{MFCC}, где M — количество блоков, на которые разбивается сигнал, а N_{MFCC} — количество рассчитываемых мел-частотных кепстральных коэффициентов для каждого блока, формирующихся на выходе описанного выше алгоритма. Установим число интервалов, в пределах которых необходимо сгруппировать значения коэффициентов, а также установим границы этих интервалов. Затем подсчитываем число попаданий значений мел-кепстральных коэффициентов в каждый интервал по всем блокам. Вместо набора векторов кепстральных коэффициентов получаем один вектор, с единой для всех образцов размерностью, которая значительно меньше размерности матрицы векторов. Размерность данного вектора можно менять исходя из необходимой точности частотного распределения (числа интервалов карманов при расчете частот), а также количества используемых кепстральных коэффициентов.

На рисунке 2 изображен вектор признаков, полученных описанным выше способом. График представляет собой, по сути, двадцать четыре последовательно расположенных гистограммы для каждого из коэффициентов. На рисунке 3а изображены четыре таких вектора для четырех разных образцов фразы «ноль, один, два», повторенных одним диктором. На рисунке 3б изображены векторы признаков для фразы «ноль, один, два», произнесенной двумя разными дикторами. Видно, что для двух разных дикторов, произносящих одну и ту же фразу в отличие от случая одного и того же диктора, значения качественно различаются.

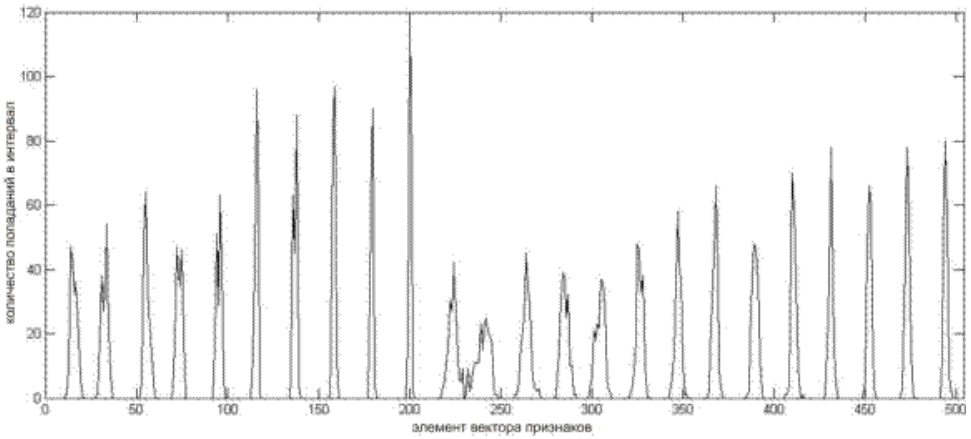

Рис. 2. Пример полученного вектора признаков для одного диктора

2. Описание модуля предварительной обработки голосовых образцов. В данной работе при проведении апробации предложенного подхода к формированию векторов признаков в качестве голосовых образцов использовались WAV-файлы с частотой дискретизации 16 кГц и разрядностью 16 бит. Использовалась база дикторов из 20 человек, по 10 повторений одной фразы. В качестве программной среды для обработки сигнала использовался пакет MATLAB с бесплатным toolbox'ом VOICEBOX [5], содержащим богатую библиотеку функций для обработки мультимедиа сигналов.

На первом этапе предобработки использовалась стандартная функция wavread, которая возвращает вектор значений амплитуд сигнала и частоту его дискретизации. Далее при помощи VAD-алгоритма (Voice Activity Detector), реализованного в функции vadsohn toolbox'a VOICEBOX, выделялись участки сигнала, не содержащие речь, и проводилось их последующее удаление.

Расчет MFCC коэффициентов был реализован путем применения функции melcepst. На вход данной функции подавались речевой сигнал S, частота его дискретизации Fs, количество кепстральных коэффи-

циентов на выходе, исключая коэффициент c_0 (в данной работе — 24), длина окна в отсчетах, на которые будет разбиваться сигнал (размер окна выбран 20 мс, что в отсчетах равно 0,02*Fs), количество фильтров в гребенке треугольных фильтров (использовано значение по умолчанию, равное примерно 2,1 на октаву) и размер перекрытия между окнами, который в данном случае равнялся половине окна. На выходе получали матрицу размером Mx24, где M — количество окон, на которые был разбит исходный сигнал. Величина зависела от длины входного сигнала.

Затем рассчитывалось частотное распределение значений для каждого из 24 коэффициентов по полученной матрице. Для первых 10 кепстральных коэффициентов были установлены границы интервала построения распределений от –10 до 10. Число интервалов было выбрано 21, т. е. от –10 до 10 с шагом 1. Для остальных кепстральных коэффициентов установим границы интервала от –2 до 2, шаг 0.2, следовательно, количество интервалов тоже 21. Таким образом, получали вектор признаков фиксированного размера: 21*24 = 504 элемента.

Далее проводилось разбиение полученных для всех голосовых образцов векторов признаков на две группы — обучающую и тестовую. В качестве образцовых

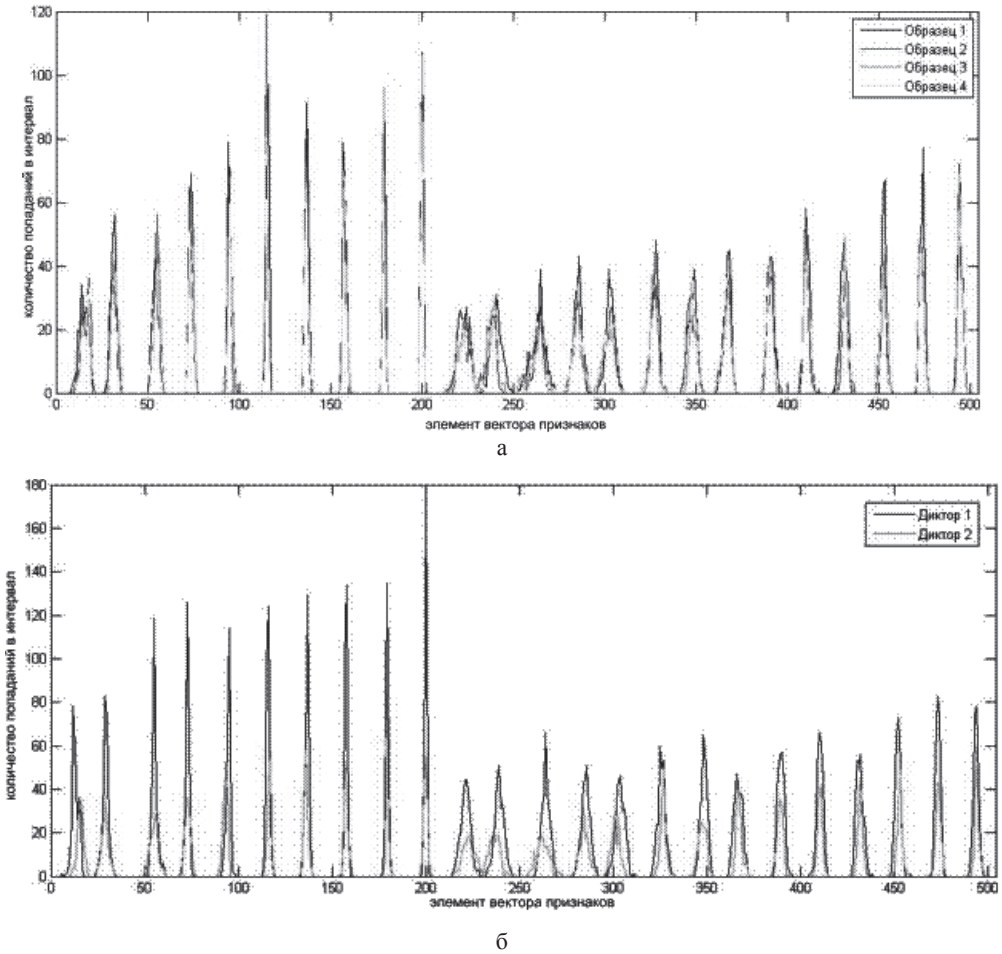

Рис. 3. Сравнение полученных векторов признаков: *а* — для одного и того же диктора; *б* — для двух разных дикторов

методов классификации применялись методы ближайших соседей [6-8] и опорных векторов [8].

3. Анализ разработанного подхода. Одними из основных характеристик методов идентификации дикторов являются ошибки первого (ложный допуск чужого) и второго (ложный недопуск своего) рода. Для оценки ошибки первого рода база обучалась на 20 дикторах, используя пять образцов для каждого диктора. Проверка проводилась по тем же дикторам, только использовались другие пять образцов. Для оценки ошибки второго рода база обучалась на 15 дикторах, используя пять образцов на диктора. Для проверки использовалось по десять образцов других пяти дикторов.

Результаты реализованного подхода сравнивались с результатами аналогичных оценок первого и второго рода для существующих методов, использующих аналогичный алгоритм. А именно: в качестве вектора признаков используется вектор из 24 мел-частотных кепстральных коэффициентов, рассчитанных для все-

го сигнала целиком, или вектор кепстральных коэффициентов, рассчитанных для фиксированного количества окон, на которые разбивается сигнал. В качестве метода классификации использовался метод опорных векторов или метод ближайших соседей.

Оценки ошибок приведены в таблице 1. Видно, что лучшие результаты достигаются при применении алгоритма, в котором вектором признаков является вектор, основанный на частотном распределении значений мел-частотных кепстральных коэффициентов, а методом классификации — метод опорных векторов. Причем наилучшие показатели ошибок первого и второго рода достигнуты при распознавании на короткой фразе, в данном случае «ноль, один, два». Такой результат может быть обусловлен тем, что одно слово — недостаточно сложная модель для распознавания диктора и происходит недообучение системы. А использование длинной фразы «ноль, один, два, три, четыре, пять, шесть, семь, восемь, девять», по всей видимости, приводит к переобучению системы.

Ошибки первого и второго рода.

№ алгоритма	Вектор признаков	Метод классификации	Ошибки первого рода, %	Ошибки второго рода, %
1	24 мел-частотных кепстральных коэффициента рассчитанные для всего сигнала целиком	Метод ближайшего соседа	4	80
2	24 мел-частотных кепстральных коэффициента рассчитанные для всего сигнала целиком	Метод опорных векторов	5	48
3	Мел-частотных кепстральныекоэффициенты рассчитанные для всего фиксированного количества окон (50 окон)	Метод ближайшего соседа	18	52
4	Мел-частотных кепстральныекоэффициенты рассчитанные для всего фиксированного количества окон (50 окон)	Метод опорных векторов	10	6
5	Вектор признаков, основанный на частотном распределении значений мел-частотныхкепстральных коэффициентов	Метод ближайшего соседа	1	54
6	Вектор признаков, основанный на частотном распределении значений мел-частотныхкепстральных коэффициентов	Метод опорных векторов	3	12

Заключение. В данной работе была предложена основанная на получении распределений мелкепстральных коэффициентов методика получения вектора признаков, характеризующих индивидуальные параметры голоса. В среде MATLAB был реализован модуль голосовой аутентификации на основе изученного метода получения вектора признаков и метода опорных векторов. Были проведены тестовые испытания разработанного модуля и показано, что такой подход к предварительной обработке акустических данных имеет хорошие характеристики по сравнению с применяемыми и может использоваться при построении эффективных систем речевой идентификации.

Библиографический список

1. Первушин Е. А. Обзор основных методов распознавания дикторов // Математические структуры и моделирование. — 2011. — Вып. 24.

2. Малинин П.В., Поляков В.В. Иерархический подход в задаче идинтификации личности по голосу с помощью проекционных методов классификации многомерных данных // Доклады Томского гос. университета систем управления и радиоэлектроники. — 2010. — № 1/1.

3. Сорокин В. Н., Вьюгин В. В., Тананыкин А. А. Распознавание личности по голосу: аналитический обзор // Информационные процессы. — 2012. — Т. 12, № 1.

4. Ganchev T., Fakotakis N., Kokkinakis G. Comparative evaluation of various MFCC implementations on the speaker verification task // 10th International Conference on Speech and Computer. — Patras, Greece, 2005.

5. VOICEBOX: Speech Processing Toolbox for MATLAB [Электронный ресурс]. — URL: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

6. Christopher M. Bishop. Pattern recognition and machine learning. — Hamburg, 2006.

7. Кучерявский С.В., Поляков В.В. Применение методов анализа многомерных данных и исследования структуры материала // Заводская лаборотория. Диагностика материалов. — 2007. — Т. 73, №8.

8. Воронцов К. В. Лекции по методу опорных векторов [Электронный ресурс]. — URL: http://www.ccas.ru/voron/download/SVM.pdf.