УДК 004.934.8'1 **DOI** 10.15622/sp.52.2

И.А. Рахманенко, Р.В. Мещеряков

АНАЛИЗ ИДЕНТИФИКАЦИОННЫХ ПРИЗНАКОВ В РЕЧЕВЫХ ДАННЫХ С ПОМОЩЬЮ GMM-UBM СИСТЕМЫ ВЕРИФИКАЦИИ ДИКТОРА

Рахманенко И.А., Мещеряков Р.В. **Анализ идентификационных признаков в речевых данных с помощью GMM-UBM системы верификации диктора.**

Аннотация. Данная статья посвящена отбору и оценке речевых признаков, используемых в задаче автоматической текстонезависимой верификации диктора. Для решения поставленной задачи была использована система верификации диктора, основанная на модели Гауссовых смесей и универсальной фоновой модели (GMM-UBM система).

Рассмотрены область применения и проблемы современных систем автоматической идентификации диктора. Произведен обзор современных методов идентификации диктора, основных речевых признаков, используемых при решении задачи идентификации диктора, а также рассмотрен процесс извлечения признаков, использованных далее. К рассмотренным признакам относятся мел-кепстральные коэффициенты (MFCC), пары линейного спектра (LSP), кепстральные коэффициенты перцептивного линейного предсказания (PLP), кратковременная энергия, формантные частоты, частота основного тона, вероятность вокализации (voicing probability), частота пересечения нуля (ZCR), джиттер и шиммер.

Произведена экспериментальная оценка GMM-UBM системы с применением различных наборов речевых признаков на речевом корпусе, включающем в себя записи 50 дикторов. Признаки отобраны с помощью генетического алгоритма и алгоритма жадного добавления-удаления.

Используя 256-компонентные Гауссовы смеси и полученный вектор из 28 признаков, была получена равная ошибка 1-го и 2-го рода (EER), составляющая 0,579 %. По сравнению со стандартным вектором, состоящим из 14 мел-кепстральных коэффициентов, ошибка EER была уменьшена на 42,1 %.

Ключевые слова: распознавание диктора, верификация диктора, Гауссовы смеси, GMM-UBM система, мел-кепстральные коэффициенты, речевые признаки, отбор признаков, обработка речи, генетический алгоритм, жадный алгоритм.

1. Введение. Задача автоматической идентификации диктора является одной из наиболее сложных задач в области обработки речи. Методы, используемые в современных системах идентификации диктора далеко не идеальны, что накладывает на такие системы определенные ограничения. Некоторые системы отлично работают в хороших акустических условиях, при минимальном уровне шума, однако заметно теряют в точности распознавания в условиях малого соотношения сигнал-шум. Требования к точности идентификации говорящего для подобных систем задают определенную планку, которая повышается с каждым годом. Повышение точности идентификации позволяет расширить область применения таких систем, включая системы биометрической многофакторной аутентификации, системы дистанционного банковского обслуживания, системы контроля доступа и многие другие. Таким обра-

зом, для удовлетворения нужд потребителей данных систем и выдвигаются высокие требования по точности распознавания диктора.

Для оценки точности идентификации диктора используют несколько характеристик, одна из которых является наиболее часто используемой — равная ошибка первого и второго рода (Equal Error Rate, EER). Равная ошибка первого и второго рода определяет ошибку распознавания диктора при условии равенства вероятности пропуска самозванца и отказа законному пользователю. Данная характеристика используется для оценки как текстозависимых, так и текстонезависимых систем идентификации диктора. Лучшие системы идентификации диктора, тестируемые на фиксированной базе данных, содержащей фразы нескольких сотен дикторов, показывают значение EER 3-5 % [1], испытания проводятся в Национальном институте стандартов и технологий США (NIST). Такая точность является недостаточной для современных систем идентификации диктора. С одной стороны, наиболее важной можно считать ошибку второго рода, когда за легального пользователя системы принимается самозванец, соответственно, можно сместить порог принятия решений системы в сторону уменьшения данной ошибки. Однако это повлечет за собой увеличение ошибок первого рода, то есть увеличит частоту отказов легальным пользователям на доступ к системе, что может повлечь за собой недовольство пользователей, использующих систему. Таким образом, необходимо уменьшение этих ошибок, то есть уменьшение EER, что позволит снизить вероятность потери конфиденциальной информации в случае применения в реальных банковских и других системах.

Задача распознавания диктора включает в себя две подзадачи: идентификацию и верификацию. Автоматическая верификация диктора — это подтверждение личности по голосу в соответствии с предъявленным им идентификатором (обычно именем данного диктора). Отличие же автоматической идентификации диктора заключается в том, что изначально неизвестен идентификатор диктора, соответственно, система должна сама определить, кем является данный диктор — законным пользователем, зарегистрированным в системе, или нарушителем (в случае решения задачи открытой идентификации) [2]. Система автоматической текстонезависимой верификации диктора, представленная в данной работе, решает задачу верификации закрытого множества дикторов, решая, присутствует ли на аудиозаписи голос заявленного диктора или нет. В данном случае, существование дикторов, не зарегистрированных в системе, не принимается во внимание.

2. Обзор современных систем идентификации диктора. К наиболее распространенным современным системам идентификации диктора по голосу можно отнести несколько видов систем: системы, осно-

ванные на Гауссовых смесях [3, 4], системы, основанные на применении i-векторов [5-11] и системы с применением нейронных сетей [6, 8, 9, 12-14].

2.1. Системы с применением комбинированного факторного анализа. Системы, основанные на применении комбинированного факторного анализа (Joint Factor Analysis, JFA), демонстрируют выдающиеся результаты в решении задачи текстонезависимой идентификации диктора [6-8].

В комбинированном факторном анализе речевой отрезок, произнесенный диктором, можно представить супервектором M, состоящим из суммы компонент, представляющих подпространство диктора и канала (сессии):

$$M = m + Vy + Ux + Dz\ , \tag{1}$$

где m — супервектор, независимый от диктора и сессии (обычно универсальная фоновая модель — от английского: universal background model, UBM), V и D задают подпространство диктора (матрицу собственных векторов голоса и диагональные остатки), и U задает подпространство сессии (матрицу собственных векторов каналов). Вектора y, x и z — зависимые от диктора и сессии факторы в соответствующих подпространствах, каждая из которых считается случайной переменной с нормальным распределением $N(0, I)$.

Применение комбинированного факторного анализа для распознавания диктора заключается в оценке подпространств (V, D, U) по размеченному соответствующим образом речевому корпусу и дальнейшей оценке факторов диктора и сессии (x, y, z), определенных по высказыванию диктора. Таким образом, удалив составляющую сессии из приведенной выше формулы, можно представить супервектор диктора как:

$$s = m + Vy + Dz\ . \tag{2}$$

Оценка соответствия диктора модели вычисляется как разность рассчитанного правдоподобия тестового высказывания диктора с компенсированной относительно сессии моделью диктора $(M - Ux)$.

Факторы канала, оцененные с помощью комбинированного факторного анализа, кроме информации о канале содержат и информацию о дикторе. На основе этого была предложена система голосовой идентификации, где используется факторный анализ для извлечения характеристик [15]. Факторный анализ задает новое пространство малой размерности, называемое пространством полной вариации. В этом

пространстве каждый речевой отрезок представлен новым вектором, называемый полным (суммарным) фактором, также этот вектор называется i-вектором. Компенсация влияния канала в этом подходе производится в данном маломерном пространстве, в отличие от многомерного пространства Гауссовых смесей.

Подход, используемый в предварительном факторном анализе, заключается в определении единого пространства вместо двух пространств дикторов и каналов. Это пространство одновременно включает в себя характеристики и диктора, и канала. В данной модели нет разделения между эффектами влияния диктора и эффектами влияния канала в Гауссовой смеси. Таким образом, новый дикторо- и каналозависимый супервектор Гауссовой смеси можно представить формулой:

$$M = m + Tw\,, \tag{3}$$

где m — дикторо- и каналонезависимый супервектор (например, универсальная фоновая модель), T — квадратная матрица малого порядка и w — случайный вектор с нормальным распределением $N(0, I)$. Компоненты вектора w являются полными факторами, а сам вектор называется вектором идентичности, или i-вектором (i-vector). Данный вектор является скрытой переменной, которая может быть задана апостериорным распределением с использованием статистики Баума — Велша.

Предположим, что имеется последовательность из L фреймов $\{y_1, y_2, ..., y_L\}$ и универсальная фоновая модель λ_{UBM}, состоящая из C компонентов смеси, заданных в пространстве признаков размерностью F. Тогда статистику Баума — Велша, необходимую для вычисления i-вектора, можно получить как:

$$N_C = \sum_{t=1}^{L} P(c \mid y_t, \lambda_{UBM});\tag{4}$$

$$F_C = \sum_{t=1}^{L} P(c \mid y_t, \lambda_{UBM}) y_t\,,\tag{5}$$

где $c = 1, ..., C$ — это индекс компонента Гауссовой смеси, и $P(c \mid y_t, \lambda_{UBM})$ соответствует апостериорной вероятности компонента смеси c, генерируемой вектором y_t. Кроме того, необходимо вычислить

централизованную статистику Баума — Велша первого порядка, основанную на математических ожиданиях универсальной фоновой модели:

$$\widetilde{F}_C = \sum_{t=1}^{L} P(c \mid y_t, \lambda_{UBM})(y_t - \mu_c), \quad (6)$$

где μ_c — математическое ожидание компонента Гауссовой смеси c. Таким образом, i-вектор для заданного речевого отрезка выбранного диктора можно вычислить по формуле:

$$w = (I + T^t \Sigma^{-1} N(u) T)^{-1} . T^t \Sigma^{-1} \widetilde{F}(u), \quad (7)$$

где $N(u)$ задается как диагональная матрица размерности $CF \times CF$, диагональные блоки которой равны $N_c I$ ($c = 1,...,C$), $\widetilde{F}_c$ — супервектор размерности $CF \times 1$, полученный объединением всей статистики Баума — Велша первого порядка $\widetilde{F}_c$ для заданного высказывания u. Σ — это диагональная матрица ковариации размерности $CF \times CF$, оцениваемая во время факторного анализа и моделирующая остаточную вариативность, не зафиксированную матрицей полной вариации T.

2.2. Системы с применением i-векторов. Используются несколько видов систем, основанных на данном представлении речи. Ряд систем построены с использованием машины опорных векторов с косинусным ядром, вычисляющая схожесть между полными факторами. Существуют системы, которые напрямую используют косинусное расстояние между факторами целевого диктора и полными факторами тестового сегмента в качестве оценки схожести. Кроме того, достаточно часто используются системы, где для вычисления оценки соответствия i-векторов, применяют вероятностный линейный дискриминантный анализ (PLDA) [5, 6, 13, 14].

Для вычисления метрики, используемой для сравнения двух i-векторов, используют несколько методов, в их числе машина опорных векторов с применением косинусного ядра и косинусное расстояние между i-векторами. Машина опорных векторов является бинарным классификатором, который пытается найти наилучший линейный разделитель между позитивными и негативными образцами. Однако можно использовать нелинейное разделение, заменив ядро машины, например на косинусное ядро.

Второй метод расчета расстояния между i-векторами — вычисление косинусного расстояния [5, 15]. При расчете косинусного расстояния между целевым и тестовым диктором результат сравнивается

с порогом, определяющим конечное решение. Преимущество данного метода — не требуется предварительное участие диктора с обучением. Разницу между двумя дикторами можно вычислять напрямую, без дополнительных вычислений и затрат, поэтому i-векторы можно рассматривать в качестве характеристик.

При использовании вероятностного линейного дискриминантного анализа в качестве основной метрики вычисляются две дополнительных матрицы — внутриклассовая матрица ковариации (WC), описывающая как отличаются i-векторы для разных речевых отрезков одного диктора, и межклассовая матрица ковариации (AC), показывающая отличия между i-векторами различных дикторов.

2.3. Системы с применением глубоких нейронных сетей. Одной из современных тенденций стало применение глубоких нейронных сетей (ГНС) в системах идентификации диктора по голосу. Глубокие нейронные сети используются как для извлечения статистик Баума — Велша, так и для извлечения новых признаков, которые формируются нейронной сетью в скрытом слое с меньшим количеством нейронов (Bottleneck Features, BNF). Возможно применение ГНС в качестве отдельного классификатора, обученного с целью идентификации диктора [16]. Кроме того, возможно применение ГНС, обученных для распознавания речи, а затем использующихся для извлечения как BNF, так и признаков, полученных из выходного слоя ГНС [6, 13, 14].

Обычно для данных целей применяют нейронные сети прямого распространения, которые намного больше (более тысячи нейронов в скрытом слое) и намного глубже (5-7 скрытых слоев) традиционных нейронных сетей. Для обучения ГНС применяют алгоритм обратного распространения ошибки и метод стохастического градиентного спуска.

Были получены результаты [14] в условиях соревнований DAC2015, в соответствии с которыми ошибка EER была уменьшена на 55% благодаря применению ГНС. В [6] применение ГНС, обученной для распознавания речи, позволило уменьшить ошибку EER на 30 % по условиям NIST SRE 2012.

3. Извлечение признаков. Одним из признаков, часто используемый в научных работах, связанных с обработкой речи и распознаванием диктора, являются мел-кепстральные коэффициенты (Mel frequency cepstral coefficients, MFCC). Мы считаем, что существуют другие признаки, которые могут содержать дополнительную информацию о дикторе, применение которой способно улучшить точность распознавания. Мы предлагаем пересмотреть набор признаков, применяемых в системах автоматической верификации диктора. Следует уделить больше внимания другим признакам, используемым в обработке

речи, таким как пары линейного спектра (line spectral pair, LSP), кепстральные коэффициенты перцептивного линейного предсказания (perceptual linear prediction cepstral coefficients — PLP), кратковременная энергия, формантные частоты, частота основного тона, вероятность вокализации (voicing probability), частота пересечения нуля (zero crossing rate, ZCR), джиттер и шиммер.

Основной набор признаков, используемый в представленной системе, — это мел-частотные кепстральные коэффициенты. Метод мел-частотного кепстрального преобразования спектра был впервые представлен в работе [17]. Мел-кепстральные коэффициенты (МКК) используются в таких областях, как распознавание диктора, распознавание речи и других задачах, связанных с обработкой речи. Наиболее часто используют 12, 13 или 14 МКК. Кроме того, часто используются дельта и двойные дельта коэффициенты, которые отражают изменения в мел-кепстральных коэффициентах во времени. В данной работе был использован вектор из 14 мел-кепстральных, 14 дельта и 14 двойных дельта коэффициентов.

Несмотря на тот факт, что в спектре речи нет признаков, которые можно было бы использовать для точного распознавания диктора, тем не менее МКК достаточно эффективно используются в задаче автоматического распознавания диктора [18]. Это возможно благодаря тому факту, что в спектре речи диктора отражается структура речевого тракта, которая позволяет отличаться голосам людей на физиологическом уровне.

Рассмотрим подробнее процесс извлечения речевых признаков (рисунок 1). На первом шаге процесса извлечения речевых признаков из аудиозаписи производится разделение ее на окна — маленькие части речевого аудио сигнала. Данные окна обрабатываются по отдельности, обработка всего сигнала целиком не производится. Длина такого окна составляет 20 мс, а смещение, по которому сигнал разбивается на окна, составляет 10 мс. После этого производится предобработка сигнала (фильтр верхних частот) и умножение на оконную функцию Хэмминга.

Для вычисления МКК после данных шагов производится дискретное преобразование Фурье (ДПФ) и переход к шкале мел. Частоты f, полученные после ДПФ, переводят к шкале мел f_{mel} с помощью преобразования:

$$f_{mel} = 1125 \ln(1 + f / 700). \quad (8)$$

Преобразование между частотами в герцах и в мелах является линейным до частоты 1000 Гц и логарифмическим выше данной ча-

стоты [19]. Для выполнения данного преобразования создается набор треугольных фильтров и вычисляется логарифм энергии в каждой полосе частот данных фильтров [17]. Последним шагом извлечения МКК является выполнение обратного ДПФ.

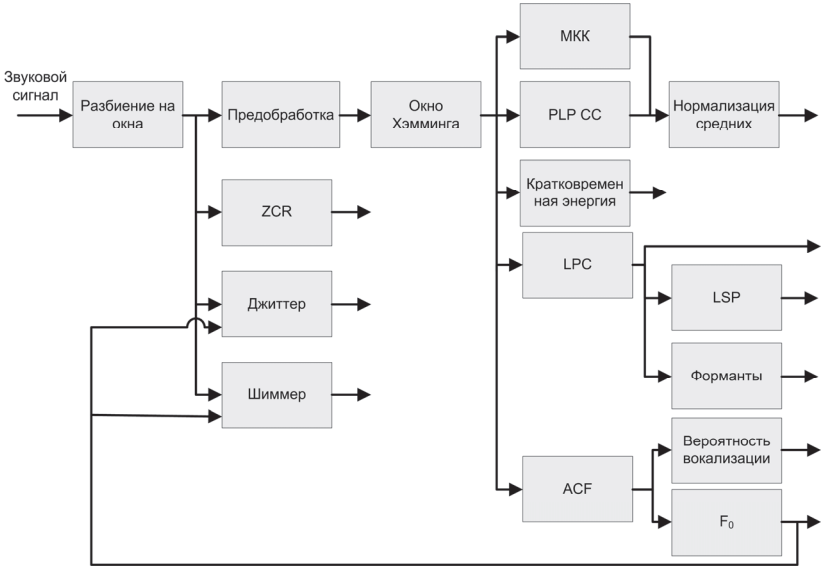

Рис. 1. Диаграмма процесса извлечения признаков

Кроме МКК к вычисляемому вектору признаков были добавлены такие признаки, как джиттер, шиммер, ZCR, кепстральные коэффициенты PLP, кратковременная энергия, вероятность вокализации (V_p), частота основного тона (F_0), форманты, LSP (рисунок 1). Вероятность вокализации вычисляется как максимум автокорреляционной функции спектра в окне. Итого, полный вектор признаков содержит 94 признака, из которых 42 — мел-кепстральные и дельта-коэффициенты. Вычисление признаков производилось с помощью библиотеки openSMILE [20].

4. GMM-UBM система. После решения вопроса выбора характеристик в системе распознавания диктора уделяют внимание методу построения решающих правил. В данной работе применяются Гауссовы смеси.

Гауссова смесь (ГС) — это параметрическая функция плотности вероятности, представленная как взвешенная сумма отдельных Гаус-

совых плотностей [21]. ГС, состоящая из C плотностей вероятности, может быть представлена формулой:

$$p(x \mid \lambda) = \sum_{i=1}^{C} w_i g(x \mid \mu_i, \Sigma_i), \quad (9)$$

где x — D -мерный непрерывный вектор данных (признаков), w_i, $i = 1,...,C$, — это вес i -го компонента смеси, и $g(x \mid \mu_i, \Sigma_i)$, $i = 1,...,C,$ это Гауссова плотность вероятности i -го компонента смеси с вектором математических ожиданий μ_i и ковариационной матрицей Σ_i. Таким образом, полную ГС можно описать множеством векторов математического ожидания, ковариационных матриц и весов смесей каждого компонента модели. ГС можно представить уравнением:

$$\lambda = \{w_i, \mu_i, \Sigma_i\}. \quad (10)$$

В итоге при решении задачи распознавания диктора каждый из дикторов представлен в системе собственной ГС λ.

Гауссовы смеси используют в задачах идентификации диктора благодаря двум наблюдениям [22]. Во-первых, было замечено, что индивидуальные компоненты смеси могут моделировать некоторое множество акустических классов. Данное множество представляет собой набор конфигураций голосового тракта диктора, что позволяет использовать их в целях идентификации. Акустические классы являются «скрытыми», так как в обучающих и контрольных данных они не размечены. Если предположить, что векторы признаков независимы друг от друга, то Гауссова смесь описывает эти классы через плотность распределения наблюдаемых векторов признаков.

Во-вторых, линейная комбинация нормальных распределений может представлять большое множество распределений акустических признаков. Достоинством Гауссовой смеси является способность точной аппроксимации распределений произвольной формы. Можно сказать, что Гауссова смесь представляет собой нечто среднее между методом векторного квантования, где распределение признаков представлено дискретным множеством шаблонов, и одним Гауссовым распределением с единственным вектором математических ожиданий и ковариационной матрицей.

Универсальная фоновая модель (УФМ, UBM) — это ГС, обученная на большом наборе речевого материала, взятого от большого

множества дикторов, ожидаемых системой во время распознавания. Благодаря этому можно использовать УФМ для проверки альтернативной гипотезы, то есть того случая, когда на записи отсутствует голос заданного диктора. Как и в [23], параметры для УФМ были обучены с помощью EM-алгоритма, а для обучения моделей дикторов была использована форма Байесовой адаптации.

Была использована ГС, состоящая из 256 компонентов, так как было замечено, что EER не уменьшался при увеличении компонент смеси. Модели дикторов были получены с помощью MAP адаптации, с адаптацией только векторов математических ожиданий и фактором релевантности $r = 10$. GMM-UBM система, описанная в текущем разделе, была создана с применением библиотеки MSR Identity Toolbox [24].

5. Экспериментальная оценка. Проведены эксперименты с применением речевого корпуса, включающего записи речи 25 дикторов-мужчин и 25 женщин. Данный речевой корпус содержит записи произнесенных без предварительной подготовки предложений, взятых из художественной литературы или поговорок [25-31]. Суммарная длина записей речи для каждого диктора составляет не менее 6 минут, включая 50 сегментов различной длины. Каждый диктор был записан на микрофон в условиях небольшого шума, частота дискретизации 8000 Гц, разрядность 16 бит.

Весь речевой корпус, состоящий из записей речи 50 дикторов, был разделен на обучающую выборку для УФМ, состоящую из записей 30 дикторов, и выборку, использующуюся для обучения и тестирования моделей дикторов, состоящую из записей оставшихся 20 дикторов. Все выборки были выполнены с равным разделением по дикторам разного пола.

Для MAP адаптации моделей дикторов использовались 40 речевых сигналов. Оставшиеся 10 сигналов каждого диктора применялись для тестирования системы верификации. В сумме было произведено 4000 тестов для каждого набора признаков, по 10 положительных (тестируется целевой диктор) и 190 отрицательных (тестируется нарушитель) для каждого диктора.

После фазы обучения, во время которой производится обучение УФМ и адаптация моделей дикторов, начинается фаза тестирования. Для каждого тестового речевого отрезка с помощью моделей ГС и УФМ была вычислена оценка верификации Λ — логарифмированное отношение правдоподобия (11). Предполагаемый диктор принимался или отвергался системой на основе порога принятия решения.

$$\Lambda(X) = \log p(X \mid \lambda_{hyp}) - \log p(X \mid \lambda_{UBM}), \qquad (11)$$

где X — тестируемый отрезок речи, λ_{hyp} — модель предполагаемого диктора, λ_{UBM} — универсальная фоновая модель. Для оценки системы голосовой верификации использовались две различных метрики: равная ошибка 1-го и 2-го рода EER и минимальная функция стоимости обнаружения (minimum detection cost function, minDCF) с параметрами SRE 2008. Функция стоимости обнаружения вычисляется как взвешенная сумма вероятности отказа целевому диктору P_{fa} и вероятности пропуска самозванца P_{miss} (12). Соответственно, минимум данной функции определяется по полученным оценкам данных вероятностей (по ошибке 1-го и 2-го рода). Результаты экспериментов отражены в таблице 1.

$$DCF = 0{,}1P_{miss} + 0{,}01P_{fa}\,. \tag{12}$$

Как видно из таблицы 1, наилучшие результаты были получены, используя вектор признаков, состоящий из 14 мел-кепстральных коэффициентов и вероятности вокализации с равной ошибкой 1-го и 2-го рода EER = 0,763 %. Минимальное значение DCF было получено с помощью набора признаков, состоящего из 14 мел-кепстральных коэффициентов, их дельт и вероятности вокализации.

Таблица 1. Результаты применения различных наборов признаков

Набор признаков	% EER	minDCF*100
MFCC+V_p	*0,763*	0,805
MFCC+Δ+V_p	1,000	*0,699*
MFCC+ Δ +ΔΔ+V_p	1,000	0,803
MFCC	1,000	0,925
MFCC+Shimmer	1,000	1,007
MFCC+Δ	1,052	0,825
MFCC+JitterDDP	1,131	1,003
MFCC+Zcr	1,131	1,031
MFCC+F_0	1,157	1,161

На рисунке 2 изображены графики кривых компромиссного определения ошибки (DET кривые) для двух наборов признаков — МКК и МКК с вероятностью вокализации. Данная кривая была получена согласно выражению 11 путем изменения порога принятия решения. Каждая точка кривой соответствует полученным ошибкам 1-го и 2-го рода при фиксированном пороге.

Было выяснено, что при добавлении некоторых признаков к стандартному вектору, состоящему из 14 МКК, результаты верификации были хуже, чем при использовании вектора только из МКК. Кроме

того, можно заметить, что при добавлении вероятности вокализации к вектору признаков, происходит уменьшение ошибки EER или уменьшение minDCF. Таким образом, можно сделать вывод, что добавление вероятности вокализации в вектор признаков улучшает эффективность работы GMM-UBM системы верификации диктора.

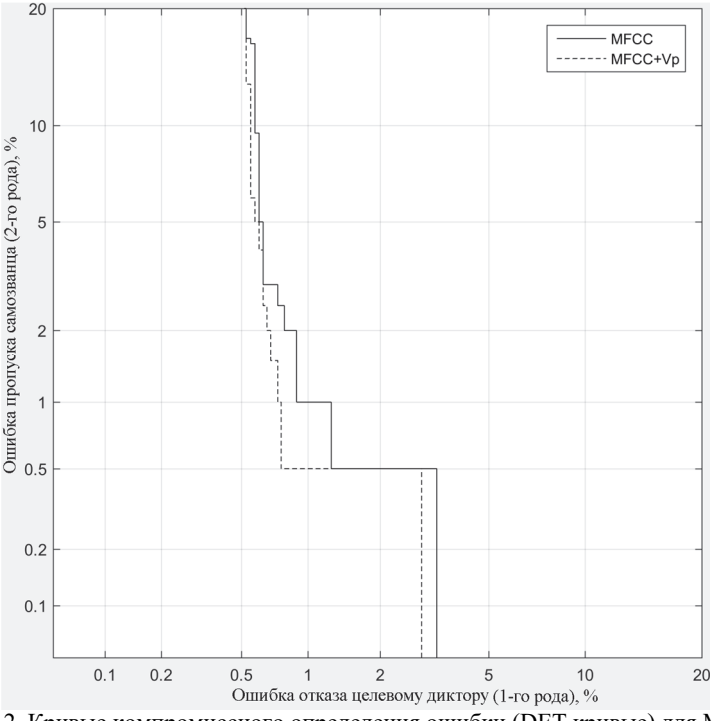

Рис. 2. Кривые компромиссного определения ошибки (DET кривые) для МКК и МКК с вероятностью вокализации

Однако полученный набор признаков нельзя считать наилучшим и дающим наименьшую ошибку EER, так как были рассмотрены только некоторые наборы признаков, выбранные авторами вручную. Выбрать наилучший набор признаков методом полного перебора в данном случае невозможно, так как для используемого количества признаков n = 94, общее количество непустых подмножеств составляет 2^n-1, что вычислительно слишком сложно. Для решения данной задачи были использованы некоторые известные методы отбора признаков, а именно метод жадного добавления-удаления (Add-del) и генетический алгоритм (ГА). Отбор признаков позволяет снизить переобучение модели и сохранить при этом наиболее информативные признаки.

Жадный алгоритм добавления-удаления признаков [32] включает в себя две жадные стратегии, то есть производится поочередное добавление и удаление признаков из текущего множества. Сначала алгоритм добавления Add последовательно добавляет признаки до тех пор, пока не начнет увеличиваться ошибка EER и еще d = 3 шагов с увеличением ошибки. После этого начинает работу алгоритм жадного удаления Del, который удаляет избыточные признаки.

Генетический алгоритм [33] осуществляет поиск наилучшего набора признаков с использованием методов естественной эволюции. Случайным образом формируется несколько наборов признаков, называемых индивидами, которые объединяются в популяцию. К полученным индивидам случайным образом применяются операции мутации и скрещивания (кроссовера), таким образом получая новые индивиды. В конце каждой итерации генетического алгоритма производится отбор лучших индивидов, для которых значение целевой функции (в данном случае EER) является наилучшим.

Несмотря на то, что генетический алгоритм позволяет достаточно быстро получить некоторый результат, его недостатками являются медленная сходимость и сложный подбор параметров.

При проведении эксперимента были использованы следующие алгоритмы: жадный алгоритм Add-del, генетический алгоритм (40 итераций), алгоритм Add-del, основанный на наилучшем индивиде, полученном по результатам ГА, и алгоритм Add-del, основанный на наилучшем наборе признаков из таблицы 1 (MFCC+V_p). Полученные результаты представлены в таблице 2. Генетический алгоритм был ограничен 40 итерациями, так как уже после третьей итерации наилучший генотип давал EER = 1 % без дальнейшего уменьшения данного показателя. При этом в последующих итерациях несколько уменьшилась функция minDCF. Это может говорить о попадании алгоритма в локальный оптимум, однако попытки перезапуска ГА не дали лучшего результата.

Таблица 2. Результаты применения алгоритмов отбора признаков

Набор признаков	Количество признаков	% EER	minDCF*100
Add-del(MFCC+V_p)	28	*0,579*	0,623
MFCC+V_p	15	0,763	0,805
ГА	50	1,000	*0,539*
ГА, Add-del	37	1,000	0,593
MFCC (базовый)	14	1,000	0,925
Add-del	22	2,079	1,827

Из таблицы видно, что наилучший результат был получен для набора из 28 признаков, полученных методом жадного добавления-удаления, для которого за основу был взят вектор из 14 мел-

кепстральных коэффициентов и вероятности вокализации. Из данного вектора был исключен 12-й мел-кесптральный коэффициент, а также добавлены 10 дельта и 2 двойных дельта мел-кестральных коэффициентов, а также 1 коэффициент линейного предсказания и 1 коэффициент LSP. Графики кривых компромиссного определения ошибки для двух наборов признаков — МКК и полученного набора из 28 признаков представлены на рисунке 3.

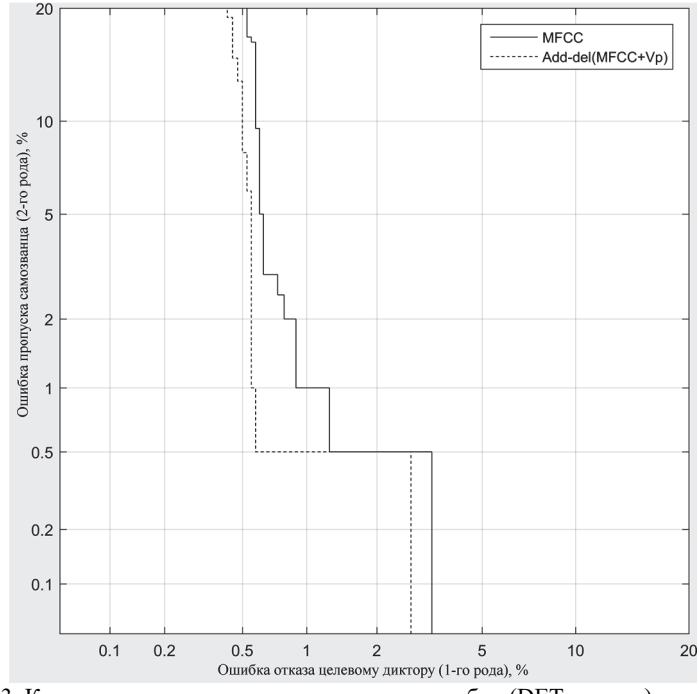

Рис. 3. Кривые компромиссного определения ошибки (DET кривые) для МКК и набора признаков, полученного с помощью жадного алгоритма Add-del

6. Заключение. Была разработана система верификации диктора, основанная на Гауссовых смесях и универсальной фоновой модели. Был произведен отбор речевых признаков с помощью алгоритма жадного добавления-удаления и генетического алгоритма. Наилучшие показатели верификации были получены при использовании вектора из 28 речевых признаков, включающий в себя мел-кепстральные коэффициенты, их дельты и двойные дельты, коэффициент линейного предсказания, линейную спектральную пару и вероятность вокализации. Равная ошибка 1-го и 2-го рода EER составляет 0,579 %. По сравнению со стандартным вектором признаков, состоящим из 14 МКК,

было достигнуто уменьшение ошибки EER на 42,1 %. В дальнейшем планируется произвести оценку представленной системы на речевом корпусе с большим множеством дикторов.

Литература

1. *Сорокин В.Н., Вьюгин В.В., Тананыкин А.А.* Распознавание личности по голосу: аналитический обзор // Информационные процессы. 2012. Т. 12. № 1. С. 1–30.
2. *Campbell Jr.J.P.* Speaker recognition: a tutorial // Proceedings of the IEEE. 1997. vol. 85. no. 9. pp. 1437–1462.
3. *Sahidullah M., Saha G.* A novel windowing technique for efficient computation of MFCC for speaker recognition // IEEE signal processing letters. 2013. vol. 20. no. 2. pp. 149–152.
4. *Motlicek P. et al.* Employment of subspace gaussian mixture models in speaker recognition // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4445–4449.
5. *Greenberg C. S. et al.* The NIST 2014 speaker recognition i-vector machine learning challenge // Odyssey: The Speaker and Language Recognition Workshop. 2014. pp. 224–230.
6. *Lei Y. et al.* A novel scheme for speaker recognition using a phonetically-aware deep neural network // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 1695–1699.
7. *Shum S.H. et al.* Unsupervised clustering approaches for domain adaptation in speaker recognition systems // Odyssey: The Speaker and Language Recognition Workshop. 2014. pp. 265–272.
8. *Stafylakis T. et al.* Compensation for phonetic nuisance variability in speaker recognition using DNNs // Odyssey: The Speaker and Language Recognition Workshop. 2016. pp. 340–345.
9. *Kenny P. et al*. Deep neural networks for extracting baum-welch statistics for speaker recognition // Proc. Odyssey. 2014. pp. 293–298.
10. *van Leeuwen D.A., Saeidi R.* Knowing the non-target speakers: The effect of the i-vector population for PLDA training in speaker recognition // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013. pp. 6778–6782.
11. *Xu L. et al.* Rapid Computation of I-vector // Odyssey: The Speaker and Language Recognition Workshop. 2016. pp. 47–52.
12. *Ahmad K.S. et al.* A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network // Advances in Pattern Recognition (ICAPR). 2015. pp. 1–6.
13. *McLaren M., Ferrer L., Lawson A.* Exploring the role of phonetic bottleneck features for speaker and language recognition // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5575–5579.
14. *Richardson F., Reynolds D., Dehak N.* Deep neural network approaches to speaker and language recognition // IEEE Signal Processing Letters. 2015. vol. 22. no. 10. pp. 1671–1675.
15. *Dehak N. et al.* Front-end factor analysis for speaker verification // IEEE Transactions on Audio, Speech, and Language Processing. 2011. vol. 19. no. 4. pp. 788–798.
16. *Variani E. et al.* Deep neural networks for small footprint text-dependent speaker verification // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 4052–4056.
17. *Davis S.B., Mermelstein P.* Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences // IEEE Transactions on Acoustics, Speech and Signal Processing. 1980. vol. 28. no. 4. pp. 357–366.
18. *Atal B.S.* Automatic recognition of speakers from their voices // Proceedings of the IEEE. 1976. vol. 64 no. 4 pp. 460–475.
19. *Jurafsky D., Martin J.H.* Speech and Language Processing: second edition // Pearson Education. New Jersey. 2009. 950 p.
20. *Eyben F., Weninger F., Gross F., Schuller B.* Recent developments in opensmile, the munich open-source multimedia feature extractor // Proceedings of the 21st ACM international conference on Multimedia. 2013. pp. 835–838.

21. *Reynolds D.A.* Gaussian mixture models // Encyclopedia of biometric recognition. Springer. Heidelberg. 2008. vol. 10. Issue 1-3. pp. 19–41.
22. *Reynolds D.A., Rose R.C.* Robust text-independent speaker identification using Gaussian mixture speaker models // IEEE Transactions on Speech and Audio Processing. 1995. vol. 3 no. 1. pp. 72–83.
23. *Reynolds D.A., Quatieri T.F., Dunn R.B.* Speaker verification using adapted Gaussian mixture models // Digital Signal Processing. 2000. vol. 10. no. 1. pp. 19–41.
24. *Sadjadi S.O., Slaney M., Heck L.* MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research // Speech and Language Processing Technical Committee Newsletter. 2013.
25. *Вольф Д.А., Мещеряков Р.В.* Модель процесса сингулярного оценивания частоты основного тона речевого сигнала // Акустический журнал. 2016. Т. 62. № 2. С. 216–226.
26. *Мещеряков Р.В., Конев А.А.* К вопросу об исследовании биологических параметров человека в защищенных системах //Доклады Томского государственного университета систем управления и радиоэлектроники. 2010. Т. 21. № 1-1. С. 131–133.
27. *Вольф Д.А., Мещеряков Р.В.* Модель и программная реализация сингулярного оценивания частоты основного тона речевого сигнала // Труды СПИИРАН. 2015. Вып. 6(43). С. 191–209.
28. *Ronzhin A.L., Karpov A.A.* Russian voice interface. Pattern Recognition and Image Analysis. 2007. vol. 17(2). pp. 321–336.
29. *Karpov A. et al.* Audio-Visual Speech Asynchrony Modeling in a Talking Head // In Proc. 10-th International Conference Interspeech. Brighton. UK. 2009. pp. 2911–2914.
30. *Ручай А.Н.* Улучшение надежности текстозависимой верификации диктора на основе формантного метода с помощью нового метода сегментации речевого сигнала // Доклады Томского государственного университета систем управления и радиоэлектроники. 2011. Т. 24. № 2-2. С. 241–246.
31. *Елистратов С.А. и др.* Сравнение параметров для выделения вокализованных сегментов и классификации гласных фонем // Доклады Томского государственного университета систем управления и радиоэлектроники. 2012. Т. 24. № 1-2. С. 171–174.
32. *Кормен Т. и др.* Алгоритмы. Построение и анализ. Глава 16. Жадные алгоритмы: пер. с англ. // Издательский дом Вильямс. 2012. 1296 с.
33. *Holland J.H.* Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence // MIT press. 1992. 232 p.

Рахманенко Иван Андреевич — ассистент кафедры безопасности информационных систем, Томский государственный университет систем управления и радиоэлектроники (ТУСУР). Область научных интересов: идентификация диктора, обработка речи, машинное обучение, программно-аппаратные средства защиты информации. Число научных публикаций — 12. ria@keva.tusur.ru; пр. Ленина, 40, Томск, 634050; р.т.: +7(3822)70-15-29.

Мещеряков Роман Валерьевич — д-р техн. наук, доцент, заведующий кафедрой безопасности информационных систем, Томский государственный университет систем управления и радиоэлектроники (ТУСУР). Область научных интересов: системный анализ, информационная безопасность, вопросы обработки информации в интеллектуальных системах, информационно-безопасные системы, идентификация диктора, обработка речи, машинное обучение, программно-аппаратные средства защиты информации. Число научных публикаций — 247. mrv@ieee.org; пр. Ленина, 40, Томск, 634050; р.т.: +7(3822)900111, Факс: +7 (3822) 900-111.

Поддержка исследований. Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации в рамках мероприятия 1.3 ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014—2020 годы» (соглашение о предоставлении субсидии № 14.577.21.0172 от 27 октября 2015 г.; уникальный идентификатор RFMEFI57715X0172).

DOI 10.15622/sp.52.2

I.A. RAKHMANENKO, R.V. MESHCHERYAKOV

IDENTIFICATION FEATURES ANALYSIS IN SPEECH DATA USING GMM-UBM SPEAKER VERIFICATION SYSTEM

Rakhmanenko I.A., Meshcheryakov R.V. **Identification Features Analysis in Speech Data Using Gmm-Ubm Speaker Verification System.**

Abstract. This paper is devoted to feature selection and evaluation in an automatic text-independent speaker verification task. In order to solve this problem a speaker verification system based on the Gaussian mixture model and the universal background model (GMM-UBM system) was used.

The application sphere and challenges of modern systems of automatic speaker identification were considered. Overview of the modern speaker recognition methods and main speech features used in speaker identification is provided. Features extraction process used in this article was examined. Reviewed speech features were used for speaker verification including mel-cepstral coefficients (MFCC), line spectral pairs (LSP), perceptual linear prediction cepstral coefficients (PLP), short-term energy, formant frequencies, fundamental frequency, voicing probability, zero crossing rate (ZCR), jitter and shimmer.

The experimental evaluation of the GMM-UBM system using different speech features was conducted on a 50 speaker set and a result is presented. Feature selection was done using the genetic algorithm and the greedy adding and deleting algorithm.

Equal error rate (EER) equals 0,579 % when using 256 component Gaussian mixture model and the obtained feature vector. Comparing to standard 14 MFCC vector, 42,1 % of EER improvement was acquired.

Keywords: speaker recognition, speaker verification, Gaussian mixture model, GMM-UBM system, mel frequency cepstral coefficients, speech features, feature selection, speech processing, genetic algorithm, greedy algorithm.

Rakhmanenko Ivan Andreevich — assistant professor of information systems security, Tomsk State University of Control Systems and Radioelectronics (TUSUR). Research interests: speaker recognition, speech processing, machine learning, hardware-software data protection solutions. The number of publications — 12. ria@keva.tusur.ru; 40, Lenin-avenue Tomsk, 634050, Russia; office phone: +7(3822)70-15-29.

Meshcheryakov Roman Valerievich — Ph.D., Dr. Sci., professor, head of information systems security, Tomsk State University of Control Systems and Radioelectronics (TUSUR). Research interests: speech analysis, speech recognition, medical technology, information security, speaker recognition, speech processing, machine learning, hardware-software data protection solutions. The number of publications — 247. mrv@ieee.org; 40, Lenin-avenue Tomsk, 634050, Russia; office phone: +7(3822)900111, Fax: +7 (3822) 900-111.

Acknowledgements. This research is supported by the Ministry of Education and Science of the Russian Federation within the limits of the project part of the state assignment of TUSUR in 2017 and 2019 (project 2.3583.2017).

References

1. Sorokin V.N., V'jugin V.V., Tananykin A.A. [Speaker Recognition: Analytical Review]. *Informacionnye process – Information Processes*. 2012. vol. 12. no. 1. pp. 1–30. (In Russ.).

2. Campbell Jr.J.P. Speaker recognition: a tutorial. Proceedings of the IEEE. 1997. vol. 85. no. 9. pp. 1437–1462.
3. Sahidullah M., Saha G. A novel windowing technique for efficient computation of MFCC for speaker recognition. IEEE signal processing letters. 2013. vol. 20. no. 2. pp. 149–152.
4. Motlicek P. et al. Employment of subspace gaussian mixture models in speaker recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4445–4449.
5. Greenberg C.S. et al. The NIST 2014 speaker recognition i-vector machine learning challenge. Odyssey: The Speaker and Language Recognition Workshop. 2014. pp. 224–230.
6. Lei Y. et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 1695–1699.
7. Shum S.H. et al. Unsupervised clustering approaches for domain adaptation in speaker recognition systems. Odyssey: The Speaker and Language Recognition Workshop. 2014. pp. 265–272.
8. Stafylakis T. et al. Compensation for phonetic nuisance variability in speaker recognition using DNNs. Odyssey: The Speaker and Language Recognition Workshop. 2016. pp. 340–345.
9. Kenny P. et al. Deep neural networks for extracting baum-welch statistics for speaker recognition. Proc. Odyssey. 2014. pp. 293–298.
10. van Leeuwen D.A., Saeidi R. Knowing the non-target speakers: The effect of the i-vector population for PLDA training in speaker recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013. pp. 6778–6782.
11. Xu L. et al. Rapid Computation of I-vector. Odyssey: The Speaker and Language Recognition Workshop. 2016. pp. 47–52.
12. Ahmad K.S. et al. A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network. Advances in Pattern Recognition (ICAPR). 2015. pp. 1–6.
13. McLaren M., Ferrer L., Lawson A. Exploring the role of phonetic bottleneck features for speaker and language recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5575–5579.
14. Richardson F., Reynolds D., Dehak N. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*. 2015. vol. 22. no. 10. pp. 1671–1675.
15. Dehak N. et al. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*. 2011. vol. 19. no. 4. pp. 788–798.
16. Variani E. et al. Deep neural networks for small footprint text-dependent speaker verification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 4052–4056.
17. Davis S.B., Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*. 1980. vol. 28. no. 4. pp. 357–366.
18. Atal B.S. Automatic recognition of speakers from their voices. Proceedings of the IEEE. 1976. vol. 64 no. 4 pp. 460–475.
19. Jurafsky D., Martin J.H. Speech and Language Processing: second edition. Pearson Education. New Jersey. 2009. 950 p.
20. Eyben F., Weninger F., Gross F., Schuller B. Recent developments in opensmile, the munich open-source multimedia feature extractor. Proceedings of the 21st ACM international conference on Multimedia. 2013. pp. 835–838.

21. Reynolds D.A. Gaussian mixture models. Encyclopedia of biometric recognition. Springer. Heidelberg. 2008. vol. 10. Issue 1-3. pp. 19–41.
22. Reynolds D.A., Rose R.C. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*. 1995. vol. 3. no. 1. pp. 72–83.
23. Reynolds D.A., Quatieri T.F., Dunn R.B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*. 2000. vol. 10. no. 1. pp. 19–41.
24. Sadjadi S.O., Slaney M., Heck L. MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research. Speech and Language Processing Technical Committee Newsletter. 2013.
25. Wolf D.A., Meshcheryakov R.V. [Process Model of singular estimation of the pitch frequency of the speech signal]. *Akusticheskij zhurnal – Acoustic magazine*. 2016. vol. 62. no 2. pp. 216–226. (In Russ.).
26. Meshcheryakov R.V., Konev A.A. [To the question of human biological parameters research in secure systems]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravlenija i radiojelektroniki – Tomsk state university of control systems and radioelectronics reports*. 2010. vol. 21. no. 1-1. pp. 131–133. (In Russ.).
27. Volf D.A., Meshcheryakov R.V. [Software Implementation of a Singular Meter of the Pitch Frequency of a Speech Signal]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2015. vol. 6(43). pp. 191–209. (In Russ.).
28. Ronzhin A.L., Karpov A.A. Russian voice interface. *Pattern Recognition and Image Analysis*. 2007. vol. 17(2). pp. 321–336.
29. Karpov A. et al. Audio-Visual Speech Asynchrony Modeling in a Talking Head. In Proc. 10-th International Conference Interspeech. Brighton. UK. 2009. pp. 2911–2914.
30. Ruchaj A.N. [Text-dependent speaker verification reliability improvement based on novel speech segmentation method]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravlenija i radiojelektroniki – Tomsk state university of control systems and radioelectronics reports*. 2011. vol. 24. no. 2-2. pp. 241–246. (In Russ.).
31. Elistratov S.A. et al. [Parameters compare for vocalized segmentation and vowel phonemes classification]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravlenija i radiojelektroniki – Tomsk state university of control systems and radioelectronics reports*. 2012. vol. 24. no. 1-2. pp. 171–174. (In Russ.).
32. Cormen T et. al. Introduction to Altghorithms. Chapter 16. Greedy Algorithms. MIT press, 2009. 1312 p. (Russ. ed.: Cormen T et. al. *Algoritmy. Postroenie i analiz. Glava 16. Zhadnye algoritmy.* Izdatel'skij dom Vil'jams. Publ. 2012. 1296 p.).
33. Holland J.H. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press. 1992. 232 p.