УДК 004.522

Г. П. Можаров, Р. С. Чеботарев

ТЕКСТОНЕЗАВИСИМЫЙ МЕТОД ИДЕНТИФИКАЦИИ ЧЕЛОВЕКА ПО ЕГО ГОЛОСУ

Рассматрен языко- и текстонезависимый методы голосовой идентификации человека, основанные на выделении из речевого сигнала признаков, характеризующих линейное предсказание поведения автокорреляционной функции кепстра голосового сигнала. На основе вектора признаков построена голосовая модель человека в виде максимально-правдоподобной смеси гауссианов, описывающих вектор признаков. Голосовая идентификация выполнена путем выбора модели, имеющей максимальную апостериорную вероятность ее восстановления по входному голосовому сигналу.

E-mail: chebotarev.roman@gmail.com

Ключевые слова: *голосовая идентификация, верификация, голосовые признаки, кепстр голосового сигнала, модель гауссовых смесей.*

Обработка голоса и речевые технологии являются сегодня одними из самых популярных направлений исследований. Повышенный интерес к этой области обусловлен большим спросом на результаты разработок систем речевого анализа, которые имеют самый широкий круг применения — от криминалистики и задач обеспечения безопасности до простых продуктов ежедневного пользования. Попытка классификации существующих речевых технологий приведена на рис. 1.

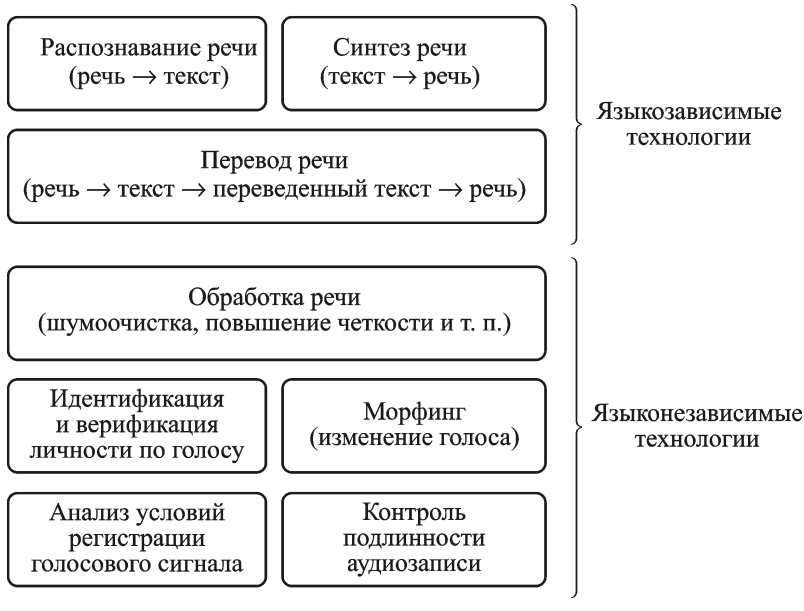

Рис. 1. Существующие речевые технологии

Одной из наиболее актуальных задач является создание технологий идентификации и верификации личности по голосу. Такие технологии могут быть использованы для построения системы контроля физического доступа на определенную территорию, систем контроля доступа к защищенной информации, для криминалистических исследований (контроль телефонного трафика операторов связи и автоматическое обнаружение представляющего интерес лица по голосу).

Применение технологии в совокупности с другими методами обработки речевой информации возможно, например, для решения задачи автоматизации работы call-центров (учет звонков с привязкой к клиентской базе, автоматический анализ и статистика запросов).

В настоящей статье рассмотрен разработанный автором языко- и текстонезависимый метод голосовой идентификации личности, а также проведен анализ точности его работы на большом количестве реальных голосовых данных, различающихся языками, условиями регистрации сигналов и гендерной принадлежностью личности.

Отличительными особенностями метода по сравнению с аналогичными технологиями западных разработчиков являются невысокие требования к качеству голосового сигнала и умеренная зависимость точности идентификации личности от условий регистрации голосового сигнала, которая варьируется в пределах $\pm 5\,\%$ при широком изменении условий регистрации. Для сравнения, точность большинства современных технологий голосовой идентификации варьируется в пределах $\pm(10–15)\,\%$ при аналогичном изменении условий регистрации голосового сигнала [1].

Под условиями регистрации голосового сигнала понимается совокупность устройства регистрации сигнала, акустической обстановки и формата хранения голосового сигнала.

Обзор существующих методов. Современные системы голосовой идентификации и верификации работают в двух режимах.

Режим обучения. Выделяются характерные признаки голоса человека, формируется его голосовая модель (голосовой отпечаток) на основе этих признаков и выполняется сохранение модели в базе данных.

Рабочий режим. Выделяются характерные признаки голосового сигнала человека и выполняется поиск в базе данных голосовой модели, соответствующей этим признакам (идентификация личности), или проверка принадлежности признаков к конкретной заданной голосовой модели (верификация личности).

Функциональная схема работы таких систем представлена на рис. 2.

Кроме этого, в режиме обучения также составляется так называемая универсальная фоновая модель (Universal Background Model,

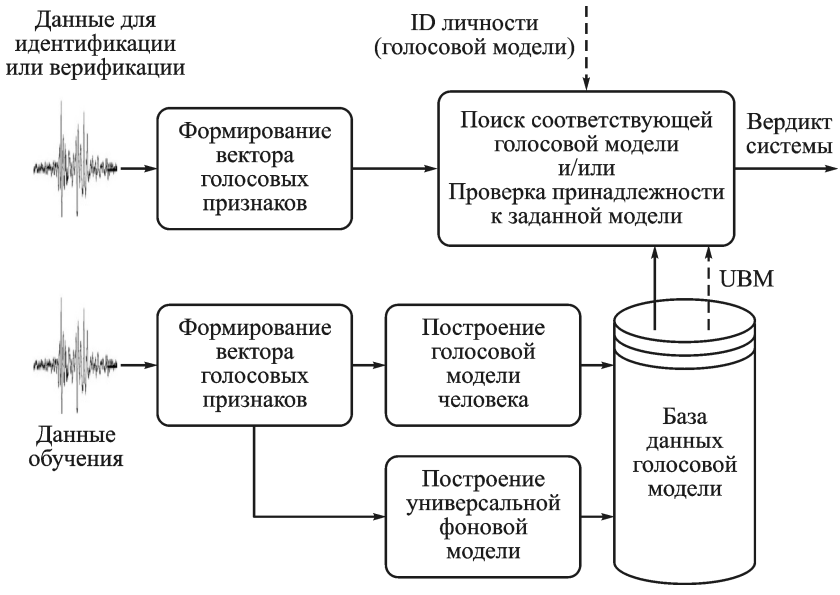

Рис. 2. Функциональная схема системы голосовой идентификации/верификации личности

UBM), которая описывает некоторые усредненные голосовые характеристики всех людей, находящиеся в базе. В рабочем режиме на основании этой модели проводится вычисление степени уникальности голосового сигнала, которая позволяет судить о достоверности идентификации/верификации и является частью аппарата принятия конечного решения.

Наибольший интерес для исследований представляют языко- и текстонезависимые методы идентификации личности. Метод является языконезависимым, если он инвариантен к языку, на котором произносится речь для целей идентификации. Аналогично, метод называется текстонезависимым, если в процессе своей работы он не получает информации о том, какую именно фразу (или слово) будет произносить человек.

В настоящее время наиболее результативным подходом к решению задач языко- и текстонезависимой идентификации личности является построение голосовых моделей на основе моделей гауссовых смесей (Gaussian Mixture Model, GMM) [2, 3]. Сами модели, как уже было отмечено, строятся на основе некоторого набора голосовых признаков, формирование которых собственно и представляет основную сложность. Наиболее распространенным методом построения голосовых признаков является формирование вектора мел-частотных кепстральных коэффициентов (Mel-Frequency Cepstral Coefficient, MFCC) из голосовой записи [1, 2].

Однако, несмотря на достаточно хорошие результаты работы в лабораторных условиях, методика GMM-MFCC не может быть исполь-

зована для построения реальных систем голосовой верификации и идентификации. Причиной тому служат очень высокие требования к качеству голосового сигнала и сильная зависимость результатов от вида обучающего материала (на основе которого составляется база голосовых моделей и фоновая модель), и условий регистрации голосового сигнала. Также недостатком являются относительно большие временные затраты на формирование вектора голосовых признаков [1–4].

Таким образом, в настоящее время существует потребность в качественном методе выделения голосовых признаков человека, способном работать с голосовыми материалами среднего качества (например запись телефонного разговора) и менее чувствительном к изменению условий регистрации голосового сигнала.

Предлагаемый метод. Суть метода голосовой идентификации заключается в использовании разработанных автором способов выделения вектора голосовых признаков и построения на его основе модели голоса человека. Вектор голосовых признаков представляет собой вектор из 12 первых коэффициентов линейного предсказания поведения автокорреляционной функции кепстра голосового сигнала.

Вычислению кепстра предшествует специфическая фильтрация голосового сигнала в диапазоне высоты звука (обычно фильтрация звука осуществляется в частотном диапазоне), которая позволяет отсечь элементы частотного разложения, слабо влияющие на голосовые характеристики, и, наоборот, подчеркнуть области, содержащие наиболее важную информацию, характеризующую индивидуальные голосовые особенности диктора.

На основании полученных векторов-признаков строятся голосовые модели путем выбора максимально-правдоподобной 1024-компонентной GMM, а также универсальная фоновая 1024-компонентная модель (UBM).

Идентификация личности (выбор голосовой модели, наиболее соответствующей заданному голосовому сигналу) осуществляется методом максимизации апостериорной вероятности. Верификация представляется как задача бинарной классификации и выполняется путем одновременной проверки гипотез принадлежности голосового сигнала к заданной голосовой модели и отсутствия его принадлежности к универсальной фоновой модели.

В реализации метода используются голосовые сигналы с частотой дискретизации $f_D = 8\,000$ Гц (сопоставимо с качеством записи мобильного телефона) и максимальной длительностью 20 с.

Формирование вектора признаков. Как уже было отмечено, вектор голосовых признаков строится из 12 первых коэффициентов

Рис. 3. Этапы формирования вектора признаков

линейного предсказания поведения автокорреляционной функции кепстра голосового сигнала. Построению кепстра предшествует фильтрация голосового сигнала в диапазоне высоты звука.

Формирование вектора голосовых признаков осуществляется по следующему алгоритму (рис. 3).

1. При необходимости исходный голосовой сигнал ограничивается по длительности (20 с) и приводится к частоте дискретизации $f_D = 8\,000$ Гц.

2. Проводится быстрое преобразование Фурье исходного сигнала, и вычисляются квадраты спектральных коэффициентов $s^2(\omega)$.

3. Частотный диапазон $[0; 0{,}5f_D]$ разбивается на 14 критических полос восприятия звука, которые соответствуют равномерному разбиению диапазона высоты звука (z, барк), получаемой из частотной шкалы (ω, Гц) по формуле

$$z = 6 \log \left(\frac{\omega}{600} + \sqrt{\left(\frac{\omega}{600} \right)^2 + 1} \right).$$

Затем определяются спектральные энергетические траектории $\ln s^2(z)$ во все критических полосах.

4. Выполняется фильтрация траекторий $\ln s^2(z)$ с целью отсечь спектральные компоненты, скорость изменения которых отлична от скорости изменения соответствующих компонентов речи, и растяжения амплитуд спектральных коэффициентов, содержащих наиболее выраженные голосовые признаки. Разработанный в процессе исследований фильтр имеет дискретную передаточную функцию вида

$$\Phi(z) = 0{,}1z^4 \frac{1 + z^{-1} - 3z^{-3} - 2z^{-4}}{1 - 0{,}9z^{-1}}.$$

5. Энергетический спектр $\ln s^2(z)$ “склеивается” из 14 критических полос и возвращается в линейный частотный масштаб $\ln s^2(\omega)$.

6. Выполняется обратное быстрое преобразования Фурье энергетического спектра, в результате которого получается кепстр $C_s(q)$, характеризующий частотно-энергетические особенности исходного сигнала в пространстве кепстрального (зависящего от частоты) времени q.

7. Вычисляется автокорреляционная функция $R_c\,(k)$ кепстра $C_s(q)$:

$$R_c\,(k) = \sum_q M\,[C_s\,(q) \cdot C_s(q-k)],$$

где $M[\cdot]$ — операция вычисления математического ожидания. Поскольку метод предполагает использование 12 кепстральных коэффициентов, то вычислять автокорреляционную функцию можно только для $k = 1 \ldots 13$.

8. Представляя значения автокорреляционной функции $R_c\,(1) \ldots \ldots R_c\,(11)$ в виде матрицы Теплица

$$\mathrm{T} = \begin{bmatrix} R_c\,(1) & R_c\,(2) & \cdots & R_c\,(11) \\ R_c\,(2) & R_c\,(1) & \cdots & R_c\,(10) \\ \vdots & \ddots & \ddots & \vdots \\ R_c\,(11) & \cdots & R_c\,(2) & R_c\,(1) \end{bmatrix},$$

а саму задачу вычисления линейного предсказания в виде

$$\begin{bmatrix} R_c\,(1) & R_c\,(2) & \cdots & R_c\,(12) \\ R_c\,(2) & R_c\,(1) & \cdots & R_c\,(11) \\ \vdots & \ddots & \ddots & \vdots \\ R_c\,(12) & \cdots & R_c\,(2) & R_c\,(1) \end{bmatrix} \begin{bmatrix} a_2 \\ a_3 \\ \vdots \\ a_{13} \end{bmatrix} = \begin{bmatrix} -R_c\,(2) \\ -R_c\,(3) \\ \vdots \\ -R_c\,(13) \end{bmatrix},$$

имеем возможность определить коэффициенты линейного предсказания поведения автокорреляционной функции эффективным с точки зрения вычислений рекурсивным методом Левинсона–Дарбина [5].

9. Кепстральные коэффициенты линейного предсказания вычисляются через рекуррентные соотношения

$$c_1 = -a_2;$$
$$c_{n+1} = -a_n - \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k}, \quad n = 1, \ldots, 11.$$

Таким образом, получаем итоговый вектор голосовых признаков $X = \{c_1, \ldots c_{12}\}$, достаточно хорошо характеризующий индивидуальные голосовые особенности человека, обусловленные физиологией его голосового тракта, и не зависящий от конкретной речевой информации, произносимой человеком.

Формирование голосовых моделей. Для построения голосовых моделей на основе векторов голосовых признаков используется 1024-

компонентная GMM. Основная идея аппарата GMM состоит в представлении плотности распределения вектора голосовых признаков X в виде взвешенной суммы гауссовых плотностей распределения:

$$p(X) = \sum_{m=1}^{M} \alpha_m p_m(X, \mu_m, D_m),$$

где $p_m(X, \mu, D)$ — гауссова плотность распределения с математическим ожиданием μ и ковариационной матрицей D, имеющей вид

$$p_m(X, \mu, D) = \frac{1}{\sqrt{2\pi \det D}} \exp(-0{,}5(X-\mu)^{\mathrm{T}} D^{-1}(X-\mu)).$$

Фактически представление плотности $p(X)$ в виде суммы M гауссианов соответствует разбиению множества голосовых параметров на M подклассов (как уже было отмечено, в предложенном методе $M = 1024$).

Также примечательно, что для GMM не важен порядок следования друг за другом определенных голосовых сигналов, поскольку данный аппарат работает с накопленными статистиками параметров.

Задача верификации пользователя по голосу представляет собой бинарную классификацию. Формально задача представляет собой проверку двух гипотез:

H_0 — фразу Y произнес человек S;

H_1 — фразу Y произнес НЕ человек S.

Оптимальной проверкой для выбора одной из двух гипотез является отношение правдоподобия. При этом процедура принятия решения выглядит следующим образом:

$$\frac{p(Y|H_0)}{p(Y|H_1)} \begin{cases} \geq \theta \ \rightarrow \ \text{принимаем } H_0; \\ < \theta \ \rightarrow \ \text{принимаем } H_1, \end{cases}$$

где $p(Y|H)$ — функция плотности вероятности для гипотезы H, оцененная на речевом сегменте Y, а θ — порог принятия решения. Математически гипотеза H может быть определена моделью λ, которая характеризует диктора S в пространстве признаков.

Для каждого человека на основании записей его речи строится голосовая модель. Для гипотезы H_1 строится универсальная фоновая модель, характеризующая всех возможных говорящих людей во всех возможных контекстах. Данная модель обучается на большом числе голосовых данных, сбалансированных по гендерному типу, а также по оборудованию и условиям регистрации голосового сигнала.

Таким образом, GMM должны быть независимо обучены для каждого человека, т.е. для каждого человека должен быть найден набор параметров $\lambda = \{\alpha_i, \mu_i, D_i\}$, $i = 1 \ldots M$ (рис. 4). Исходными данными

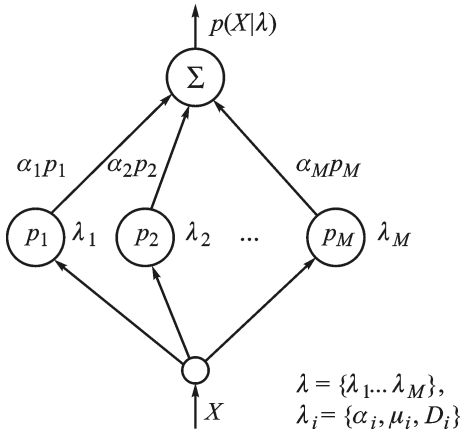

Рис. 4. Принцип построения голосовой модели на основе GMM

для обучения являются векторы голосовых признаков $X = \{c_1, \dots c_{12}\}$. Обучение GMM осуществляется по алгоритму максимального правдоподобия (Expectation-Maximization, EM). Правдоподобие модели λ на последовательности обучающих векторов $T = \{X_1, \dots X_T\}$ может быть оценено как

$$p(T \mid \lambda) = \prod_{t=1}^{T} p(X_t \mid \lambda).$$

Идея алгоритма максимального правдоподобия заключается в последовательном изменении параметров модели $\lambda_n \to \lambda_{n+1}$ таким образом, чтобы $p(T \mid \lambda_{n+1}) \geq p(T \mid \lambda_n)$ до тех пор, пока не будет достигнут порог сходимости, или пока алгоритм не будет остановлен. В рассматриваемой методике оценка максимального правдоподобия проводится по алгоритму Баума–Уэлша, который традиционно используется для нахождения неизвестных параметров скрытых марковских моделей [6] (рис. 4).

Схожим образом формируется фоновая модель λ_{UBM}, за исключением того, что последовательность обучающих векторов T составляется из всех возможных векторов голосовых признаков X.

Идентификация и верификация. Группа людей $G = \{S_1, \dots\ S_k\}$ в системе голосовой идентификации представлена своими голосовыми отпечатками в базе GMM $\Lambda = \{\lambda_1, \dots\ \lambda_k\}$. Определение, какой из моделей в базе Λ наиболее соответствует некоторый вектор признаков X, происходит путем выбора такой модели λ_m, которая имеет максимум апостериорной вероятности:

$$\hat{S} = \arg \max_{1 \leq m \leq k} \Pr(\lambda_m \mid X) = \arg \max_{1 \leq m \leq k} \frac{p(X|\lambda_m) \Pr(\lambda_m)}{p(X)},$$

или с учетом равной вероятности появления каждого человека из базы

голосовых моделей

$$\hat{S} = \arg \max_{1 \le m \le k} p\,(X \mid \lambda_m).$$

После выбора наиболее соответствующей голосовой модели λ_m решается задача верификации:

$$\frac{p(X|\lambda_m)}{p(X|\lambda_{UBM})} \begin{cases} \ge \theta \to X \text{ соответствует модели } \lambda_m; \\ < \theta \ \to X \text{ соответствует НЕ модели } \lambda_m, \end{cases}$$

где порог $\theta = 1,65$ был выбран в ходе экспериментов как оптимальный с точки зрения максимальной точности идентификации. Примечательно, что значение этого оптимального порога изменялось весьма незначительно при различных способах проведения эксперимента.

Экспериментальные результаты. Рассматриваемый метод голосовой идентификации был полностью реализован в среде Mathworks Matlab R2010. В качестве материалов для обучения и тестирования системы были использованы данные конкурсов систем распознавания дикторов NIST SRE в 2004, 2006 и 2008 гг. [7], из которых были отобраны фонограммы дикторов, имеющих по 6–10 голосовых записей длительностью около 16 с (табл. 1). Фонограммы содержат большое число разнообразных фраз, произносимых на разных языках в условиях различной акустической обстановки (помещение, улица и т.п.).

Таблица 1

Используемая база фонограмм

Гендерный состав	Общее число отобранных участников и их фонограммы	Каналы		
		телефон–телефон	микрофон–микрофон	телефон–микрофон
М	Дикторы	473	95	92 тел. + 95 мик.
	Фонограммы	3928	910	1374
Ж	Дикторы	626	122	121 тел. + 122 мик.
	Фонограммы	5153	1173	1829

Были проведены всевозможные способы обучения универсальной фоновой модели (UBM), формирование голосовых моделей всех дикторов и тестирование системы идентификации на этих данных. Фонограммы для обучения и тестирования случайно выбирались из доступных таким образом, чтобы фонограммы, использовавшиеся для построения голосовых моделей, не участвовали в тестировании.

Для обучения универсальной фоновой модели (UBM) дополнительно были выбраны фонограммы из [7], не используемые ни для формирования голосовых моделей, ни для тестирования (табл. 2).

Таблица 2

Число фонограмм для обучения UBM

Гендерный состав дикторов	Каналы		
	телефон–телефон	микрофон–микрофон	телефон–микрофон
М	642	871	604
Ж	737	1342	597

В результате эксперимента были определены вероятности достоверной идентификации личности диктора по голосовой записи при различных данных для обучения UBM и данных для тестирования. Результаты приведены в табл. 3, 4 и 5.

Таблица 3

Точность идентификации в канале телефон–телефон

Данные для обучения UBM	Данные для тестирования		
	М	Ж	М + Ж
М	96,0 %	–	–
Ж	–	95,7 %	–
М + Ж	95,2 %	94,9 %	**93,7 %**

Таблица 4

Точность идентификации в канале микрофон–микрофон

Данные для обучения UBM	Данные для тестирования		
	М	Ж	М + Ж
М	97,2 %	–	–
Ж	–	97,9 %	–
М + Ж	96,1 %	96,8 %	**94,3 %**

Таблица 5

Точность идентификации в канале телефон–микрофон

Данные для обучения UBM	Данные для тестирования		
	М	Ж	М + Ж
М	92,1 %	–	–
Ж	–	92,6 %	–
М + Ж	91,6 %	92,0 %	**90,4 %**

Как следует из табл. 3–5, метод демонстрирует достаточно высокую точность текстонезависиомой голосовой идентификации лично-

сти, сравнимую с результатами ведущих мировых разработчиков подобного рода систем [1–4]. Имеет место умеренная зависимость точности идентификации от условий регистрации голосового сигнала (как для целей обучения и составления голосовых моделей, так и непосредственно для идентификации), а также гендерного состава базы голосовых моделей.

Анализ финальных и промежуточных результатов показал, что значительное число ошибок идентификации приходится на неверный выбор голосовой модели λ_m, соответствующей вектору признаков X. Важность этого замечания обусловлена тем, что система голосовой идентификации может выдавать ложный сигнал, даже в случае построения адекватных голосовых моделей, за счет одного только неполноценного аппарата выбора конкретной голосовой модели λ_m (или уведомления об отсутствии таковой), соответствующей конкретному вектору признаков X.

В будущем планируется провести анализ влияния условий регистрации голосового сигнала на кепстральные коэффициенты в целях разработки метода формирования вектора голосовых признаков, слабо чувствительных к условиям регистрации голосового сигнала. Также планируется разработка более совершенного классификатора голосовых моделей (т.е. выбора голосовой модели, наиболее соответствующей заданному вектору голосовых признаков). В настоящее время ведется исследование применимости для этих целей аппарата машин опорных векторов, в частности, быстро обучаемых лагранжевых машин (LSVM).

Заключение. Разработан метод языко- и тескстонезависимой голосовой идентификации личности, точность работы которого сопоставима с точностью систем голосовой идентификации ведущих мировых разработчиков.

Отличительной особенностью метода является умеренная зависимость точности идентификации от условий регистрации голосового сигнала (устройства регистрации сигнала, акустическая обстановка, каналы передачи сигнала).

Данный метод может быть положен в основу работы систем голосовой идентификации и верификации как коммерческого применения, так и систем, обеспечивающих контроль физического и информационного доступа с повышенными требованиями к защищенности.

СПИСОК ЛИТЕРАТУРЫ

1. R e y n o l d s D. Experimental evaluation of features for robust speaker identification // IEEE Trans. On Speech and Audio Processing, 1994. – Vol. 2. No. 4. – P. 639–643.

2. B i m b o t F. et al. A tutorial on text-independent speaker verification // EURASIP J. on Applied Signal Processing. – 2004. – No. 4. – P. 430–451.
3. R e y n o l d s D., R o s e R. Robust text-independent speaker identification using Gaussian mixture speaker models // IEEE Trans. On Speech and Audio Processing. – 1995. – No. 3. – P. 72–83.
4. H e r m a n s k y H., M o r g a n N. RASTA processing of speech // IEEE Trans. On Speech and Audio Processing. – 1994. – Vol. 2. No. 6. – P. 578–589.
5. M u s i c u s B. Levinson and fast Choleski algorithms for Toeplitz and Almost Toeplitz Matrices // RLE TR, MIT, 1998. – No. 538.
6. W e l c h L. Hidden Markov Models and the Baum-Welch algorithm // IEEE Information Theory Society Newsletter, 2003.
7. http://www.itl.nist.gov/iad/mig/tests/sre/

Статья поступила в редакцию 15.12.2011