УДК 621.391.037.372

А.В. Козлов, О.Ю. Кудашев, Ю.Н. Матвеев,
Т.С. Пеховский, К.К. Симончик, А.К. Шулипа

СИСТЕМА ИДЕНТИФИКАЦИИ ДИКТОРОВ ПО ГОЛОСУ ДЛЯ КОНКУРСА NIST SRE 2012

Козлов А.В., Кудашев О.Ю., Матвеев Ю.Н., Пеховский Т.С., Симончик К.К., Шулипа А.К.
Система идентификации дикторов по голосу для конкурса NIST SRE 2012.
Аннотация. Приведено описание системы идентификации дикторов по голосу, разработанной для конкурса по оцениванию систем распознавания дикторов NIST SRE 2012.
Ключевые слова: биометрическая идентификация, идентификация диктора, распознавание по голосу, GMM, PLDA, JFA, NIST.

Kozlov A.V., Kudashev O.Yu., Matveev Yu.N., Pekhovsky T.S., Simonchik K.K., Shulipa A.K.
Speaker recognition system for the NIST SRE 2012.
Abstract. A description of a speaker identification system by voice is presented. This system was developed for submission to speaker recognition system evaluation at NIST SRE 2012.
Keywords: biometrics identification, speaker recognition, voice recognition, GMM, PLDA, JFA, NIST.

1. Введение. Процедура распознавания (идентификации/верификации) личности по голосу заключается в сравнении записей голоса, сделанных с помощью различных аудиоустройств, с целью выявить совпадение принадлежности голоса на этих записях одному и тому же человеку (диктору). В автоматических системах идентификации по результатам сравнения выводится оценка вероятности или иной меры сходства/различия голосов дикторов на интересующих фонограммах.

Системы идентификации дикторов по голосу относятся к классу биометрических систем, область использования которых достаточно обширна:

– автоматическая верификация клиентов при удаленном доступе по телефонному каналу;
– обработка речевых баз данных;
– криминалистические исследования.

В данной статье представляется описание текстонезависимой системы автоматической идентификации дикторов по голосу, разработанной ООО «Центр речевых технологий» (ЦРТ) для участия в международном конкурсе по оцениванию систем распознавания дикторов NIST SRE 2012. Эта система является развитием системы, представленной ЦРТ на предыдущем конкурсе 2010 года [1], и при ее создании использовался ряд новых методов и их комбинаций [2].

В профессиональной среде NIST SRE (Speaker Recognition Evaluation) называют неофициальным чемпионатом мира по голосовой идентификации. Начиная с 1996 года, этот конкурс каждые два года проводится Национальным институтом стандартов и технологий США (National Institute of Standards and Technology, NIST). Его основная цель – оценить уровень существующих технологий распознавания дикторов и определить перспективные направления развития индустрии. Регулярно в конкурсе принимают участие ведущие компании, университеты и лаборатории со всего мира. В конкурсе 2012 года участвовало 49 научных команд.

Отличительными особенностями конкурса 2012 года [4], по сравнению с конкурсами прошлых лет [5, 6], было два дополнительных условия: наличие нескольких сессий одного диктора для построения эталона и добавленного искусственного зашумления тестовых записей.

Количество фонограмм для построения эталонной модели диктора варьировалось от одной до нескольких десятков. При этом часть записей могла быть сделана в телефонном канале, другая – в микрофонном канале. Для идентификации предъявлялась единственная фонограмма, для которой заранее были известны только тип канала и пол диктора.

На рис. 1 приведена гистограмма распределения фонограмм телефонного канала связи по отношению сигнал/шум (ОСШ). Моды гистограммы соответствуют различным уровням искусственного зашумления тестовых записей: без зашумления (правая мода), с ОСШ 15 дБ (средняя мода) и ОСШ 6 дБ (левая мода).

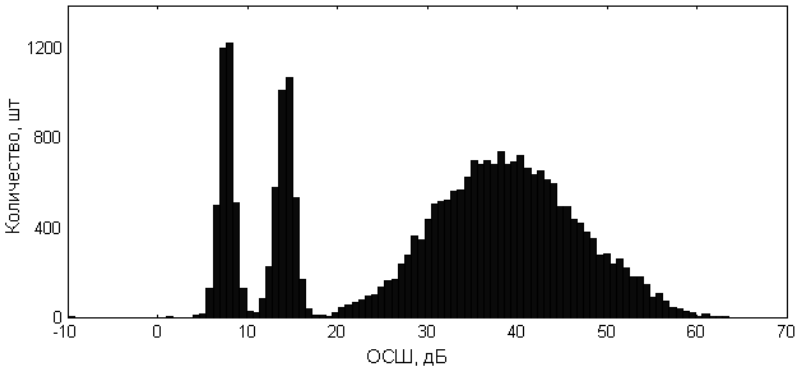

Рис. 2. Гистограмма распределения ОСШ фонограмм тестовых записей NIST SRE 2012, сделанных по телефонному каналу.

Еще одной особенностью конкурса NIST SRE 2012 являлось то, что функция ошибки детектирования (DCF) [4] представляла собой полусумму функций ошибок детектирования конкурсов NIST SRE 2008 [5] и NIST SRE 2010 [6]:

$$DCF = 0{,}5 \cdot DCF_{2008} + 0{,}5 \cdot DCF_{2010},$$

где $DCF_{2008} = FRR + 99 \cdot FAR$,

$DCF_{2010} = FRR + 999 \cdot FAR$,

FRR (*false rejection error rate*) – вероятность ошибки ложного отклонения;

FAR (*false acceptance error rate*) – вероятность ошибки ложного пропуска.

Высокая стоимость вероятности ошибки ложного отклонения означает актуальность использования автоматической идентификации по голосу, особенно для систем большого масштаба, где априорная вероятность появления целевого диктора очень мала.

2. Основные принципы работы системы идентификации личности по голосу. Работа системы автоматической идентификации личности по голосовым признакам происходит в несколько этапов [2]:

1. Выделение речевых сегментов на фонограмме.
2. Извлечение речевых признаков на речевых сегментах.
3. Построение моделей диктора как для эталонной/эталонных фонограмм, так и для тестовой фонограммы (физически модели могут храниться в оперативной или постоянной памяти, что упрощает работу с исходными данными).
4. Сравнение моделей - в результате такого сравнения определяется, насколько голос неизвестного диктора на тестовой фонограмме похож на голос известного диктора и, следовательно, принадлежат ли записи речи одному человеку или разным людям.

Принцип работы системы идентификации диктора основан на выделении речи из фонограмм и последующем попарном сравнении биометрических признаков (содержащихся в голосе индивидуальных, идентификационно значимых признаков личности).

Выделение и сравнение биометрических признаков производится с использованием различных языко- и текстонезависимых методов идентификации дикторов по голосу. Система распознавания диктора называется текстонезависимой, если она не содержит информации о том, что именно диктор будет произносить (система обучается и тестируется на произвольных речевых данных) [3].

Как уже было сказано ранее, для каждого целевого диктора строится набор моделей по каждому из методов. По каждому из методов

для спорной фонограммы проводится сравнение с эталоном, полученные оценки сравнения объединяются с помощью так называемого «обобщенного решения», результатом которого является логарифм правдоподобия сходства дикторов LLR (Log Likelihood Ratio) на фонограммах. Если полученная оценка LLR оказывается выше порога принятия решения, то спорная фонограмма кандидата приписывается целевому диктору. Если иначе, то заявленный кандидат признается самозванцем.

Представленная на конкурс NIST SRE 2012 система включала в себя модули выделения речи, извлечения речевых признаков, построения голосовых моделей, сравнения голосовых моделей.

3. Выделение речевых сегментов на фонограмме. Для выделения речевых сегментов использовались специальные алгоритмы предобработки всего сигнала [9, 7]. Так как на реальных записях, сделанных в обычных офисных или бытовых условиях, часто присутствуют посторонние сигналы, такие как импульсные и мультитональные помехи, музыка [8], а также перегруженные участки речи [10], предварительная отбраковка непригодных для анализа участков фонограммы позволяет повысить эффективность дальнейшей обработки.

Тестовые записи интервью NIST SRE 2012, сделанные в микрофонном канале, представляли собой стерео записи интервью целевого диктора (интервьюируемого) оператором. Особенность условий заключалась в том, что левый канал был записан с диктофона, расположенного «где-то в комнате», а правый – с микрофона, закрепленного около губ оператора [4]. Таким образом, ни один из каналов не содержал «чистой» речи целевого диктора.

Для выделения речи целевого диктора в микрофонных тестах SRE 2012 авторами был использован алгоритм стерео фильтрации [11], основная блок-схема которого представлена на рис. 2.

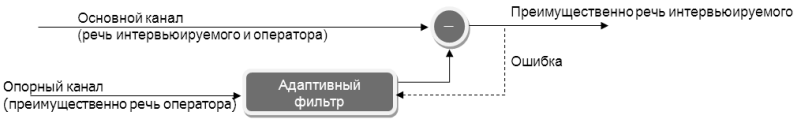

Рис. 2. Схема алгоритма стерео фильтрации для выделения речи целевого диктора

Поскольку отфильтрованный сигнал содержал преимущественно речь целевого диктора, он подавался в качестве входного на модуль выделения речи. Однако для дальнейшего построения модели речи целевого диктора отфильтрованный сигнал не использовался. Полу-

ченные с его помощью метки речевых сегментов далее применялись к левому каналу исходного сигнала, и по сегментам речи целевого диктора строилась его голосовая модель.

Телефонные тестовые записи обрабатывались модулем детектора речи напрямую, поскольку были записаны в стерео формате, каждый канал которого содержал только одного диктора.

4. Извлечение речевых признаков на речевых сегментах. В качестве речевых признаков использовались векторы мел-частотных кепстральных коэффициентов (mel-frequency cepstral coefficients, MFCC) размерностью 13 с шагом окна анализа 20 мс. Также вычислялись 13 первых производных и 13 вторых производных MFCC-коэффициентов по времени – размерность всего вектора (супервектора) составляла 39. Для каждого из векторов применялась процедура вычитания кепстрального среднего (CMS) [12].

Для повышения надежности системы в качестве дополнительных речевых признаков были использованы линейно-частотные кепстральные коэффициенты (linear-frequency cepstral coefficients, LFCC), что обеспечило повышение качества идентификации в микрофонном канале [13].

5. Построение и сравнение моделей голосов дикторов. На данный момент наиболее распространенными подходами к решению задач текстонезависимой идентификации являются подходы на основе использования моделей гауссовых смесей (Gaussian mixture models, GMM) [14].

Для построения GMM-моделей, в числе прочих, использовался метод совместного факторного анализа (Joint Factor Analysis) [15, 16, 17, 28]. Построение GMM-модели голоса в методе JFA производится путем следующего представления супервектора средних для h-й сессии s-го диктора:

$$\vec{\boldsymbol{\mu}}(s,h) = \vec{\boldsymbol{\mu}}_0 + \hat{\mathbf{U}}\vec{\mathbf{x}}(s,h) + \hat{\mathbf{V}}\vec{\mathbf{y}}(s) + \hat{\mathbf{D}}\vec{\mathbf{z}}(s)\,,$$

где $\vec{\boldsymbol{\mu}}_0$ — супервектор средних модели JFA; $\vec{\mathbf{z}}(s)$, $\vec{\mathbf{y}}(s)$ — факторы моделирования междикторской вариативности; $\vec{\mathbf{x}}(s,h)$ – низкоразмерный вектор подпространства каналов-сессий; $\hat{\mathbf{V}}$ — матрица собственных голосов, $\hat{\mathbf{U}}$ — матрица собственных каналов, $\hat{\mathbf{D}}$ — диагональная матрица.

На стадии построения JFA-модели, ее гиперпараметры $\boldsymbol{\Lambda} = \left\{\hat{\boldsymbol{\Sigma}}, \vec{\boldsymbol{\mu}}_0, \hat{\mathbf{V}}, \hat{\mathbf{U}}, \hat{\mathbf{D}}\right\}$ получают на всей базе обучения методом максимального правдоподобия с использованием EM-алгоритма. Однако, в

отличие от метода MLED (maximum likelihood eigen-decomposition) [18], обобщенный скрытый вектор $\vec{\mathbf{X}} = \{\vec{\mathbf{x}}, \vec{\mathbf{y}}, \vec{\mathbf{z}}\}^T$ находят с помощью факторного анализа в MAP-точке [20]. Здесь важно отметить, что в самом начале процедуры для каждого произнесения необходимые высокоразмерные статистики Баума–Уэлша получают на основании универсальной фоновой модели (Universal Background Model, UBM) и далее их фиксируют. Таким образом, JFA – метод на основе факторного анализа, входными данными для которого являются супервектора статистик и средних GMM-моделей, а скрытые переменные имеют нормальное распределение.

В качестве инициализации гиперпараметра $\vec{\boldsymbol{\mu}}_0$ используют супервектор средних дикторонезависимой UBM. А для инициализации диагональной матрицы $\hat{\mathbf{D}}$ используется ее значение, найденное из уравнения:

$$\hat{\mathbf{I}} = \tau \hat{\mathbf{D}}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{D}},$$

где τ — фактор релевантности в классической MAP-адаптации [18]. Матрицы $\hat{\mathbf{V}}$ и $\hat{\mathbf{U}}$ инициализируются случайным шумом.

Тестовый проход представленного метода начинается с создания JFA-модели диктора, представляющая собой каналонезависимый супервектор средних $\vec{\boldsymbol{\mu}}(s)$ для эталонной фонограммы:

$$\vec{\boldsymbol{\mu}}(s) = \vec{\boldsymbol{\mu}}_0 + \hat{\mathbf{V}}\vec{\mathbf{y}}(s) + \hat{\mathbf{D}}\vec{\mathbf{z}}(s)$$

Учитывая произнесение эталона диктора $\mathbf{O}_\mathbf{r}$ и гиперпараметры $\boldsymbol{\Lambda} = \left\{\hat{\boldsymbol{\Sigma}}, \vec{\boldsymbol{\mu}}_0, \hat{\mathbf{V}}, \hat{\mathbf{U}}, \hat{\mathbf{D}}\right\}$, мы находим этот супервектор, используя совместную MAP-оценку в точке для апостериорных значений скрытых переменных $\vec{\mathbf{x}}(s,h)$, $\vec{\mathbf{y}}(s)$, $\vec{\mathbf{z}}(s)$.

Далее рассчитывается байесовская оценка вероятности принадлежности спорного произнесения $\mathbf{O}_\mathbf{s}$ заданной JFA-модели эталонной фонограммы. Полученная оценка LLR показывает степень сходства голосов дикторов на эталонной и спорной фонограммах:

$$LLR = \frac{1}{T}\log\left[\frac{\int P(\mathbf{O}_\mathbf{s} \mid \vec{\boldsymbol{\mu}}(s) + \hat{\mathbf{U}}\vec{\mathbf{x}}(s,h))N(\vec{\mathbf{x}}(s,h) \mid 0, I)d\mathbf{x}}{\int P(\mathbf{O}_\mathbf{s} \mid \vec{\boldsymbol{\mu}}_0 + \hat{\mathbf{U}}\vec{\mathbf{x}}(s,h))N(\vec{\mathbf{x}}(s,h) \mid 0, I)d\mathbf{x}}\right],$$

где $\mathbf{O}_\mathbf{s} = \left\{\vec{\mathbf{o}}_1, ..., \vec{\mathbf{o}}_\mathbf{T}\right\}$ — совокупность векторов MFCC произнесения диктора на спорной фонограмме.

На основании таких оценок по всем парам сравнения на тестовой базе строится итоговый DET-график, по которому можно сделать оценку мощности исследуемой системы идентификации.

Значения равновероятной ошибки (equal error rate, EER) принятия чужого и отбрасывания своего диктора для метода на основе JFA зависят от длительности сравниваемых речевых фрагментов и могут достигать величины 1–2%.

В дополнение к методу классического факторного анализа, был разработан ряд методов, в которых рассматривается только одно, низкоразмерное, пространство, так называемое пространство «полной изменчивости», которое отображает как междикторскую, так и межканальную изменчивость [21]. Таким образом, при построении GMM-модели диктора учитывается только суммарное влияние факторов, что не даёт возможности выполнить, например, компенсацию эффектов канала на этом этапе (в отличие от метода факторного анализа). Для компенсации канальных искажений потребуется использование дополнительной операции на основе вероятностного линейного дискриминантного анализа (Probabilistic Linear Discriminative Analysis, PLDA) [21].

Несмотря на то, что, подобно JFA, метод PLDA является генеративным и даёт возможность произвести оценку факторов межканальной и междикторской вариативности, тем не менее, он имеет свои отличительные особенности.

Во-первых, каждое произнесение представляется в виде низкоразмерного вектора в пространстве с базисом, представленным матрицей полной изменчивости. Это выгодно отличает метод PLDA от JFA, где для каждого произнесения необходимо иметь высокоразмерные статистики Баума–Уэлша.

Во-вторых, для описания априорных распределений факторов вариативности, как для тестовых произнесений, так и произнесений из базы обучения применяются распределения Стьюдента с «тяжелыми хвостами» (heavy tailed priors), позволяющие получить устойчивые оценки параметров модели по отношению к выбросам.

В третьих, верификационная оценка на этапе принятия решения о выборе гипотезы при сравнении тестового и эталонного произнесений, производится, как показывается далее, исходя из симметричной относительно этих произнесений оценки PLDA.

В пространстве полной изменчивости супервектор (вектор математических ожиданий) GMM модели диктора имеет вид [22]:

$$\vec{\boldsymbol{\mu}}(s,h) = \vec{\boldsymbol{\mu}}_0 + \hat{\mathbf{T}}\vec{\boldsymbol{\omega}}(s,h) \ ,$$

где $\mathbf{T}$ — прямоугольная матрица из базисных векторов пространства полной изменчивости; $\vec{\omega}(s,h)$ — низкоразмерный вектор факторов полной изменчивости (в литературе имеет также название «i-вектор» [21]).

Далее задачей PLDA является декомпозиция речевых данных на дикторозависимую компоненту и компоненту, учитывающую влияние эффектов канала. Генеративная модель PLDA выглядит следующим образом:

$$\vec{\omega}(s,h) = \vec{\omega}_0 + \hat{\mathbf{U}}_1\vec{\mathbf{x}}_1(s) + \hat{\mathbf{U}}_2\vec{\mathbf{x}}_2(s,h) + \vec{\boldsymbol{\varepsilon}},$$

где $\vec{\omega}_0$ — средний вектор модели; $\hat{\mathbf{U}}_1$ — матрица, столбцы которой составляют базис междикторской вариативности или «собственные голоса» (eigenvoices); $\hat{\mathbf{U}}_2$ — матрица, столбцы которой составляют базис межканальной вариативности или «собственные каналы» (eigenchannels); $\vec{\mathbf{x}}_1(s)$ — вектор скрытых параметров модели, определяет факторы в пространстве междикторской вариативности; $\vec{\mathbf{x}}_2(s,h)$ — вектор скрытых параметров модели, определяет факторы влияния каналов записи фонограмм; $\vec{\boldsymbol{\varepsilon}}$ — вектор шумовой составляющей модели.

В настоящей работе предполагается гауссов характер априорных распределений для скрытых переменных и шума. В случае гауссова PLDA-анализатора можно получить точное решение процедуры его ML-обучения. В работе [27] П. Кенни реализовал ML-обучение PLDA с использованием так называемого VBA-inference в двух вариантах распределения скрытых переменных: нормального и Стьюдента.

После того, как были определены параметры генеративной модели PLDA, необходимо на этапе принятия решения произвести вероятностную оценку гипотезы о принадлежности тестового произнесения к эталонному.

Байесовская оценка верификации для эталонного $\vec{\omega}_1$ и тестового $\vec{\omega}_2$ выгодно отличается от LLR-оценки в методе JFA своей симметрией:

$$Score = \log\frac{P(\vec{\omega}_1,\vec{\omega}_2 \mid T)}{P(\vec{\omega}_1 \mid I)\cdot P(\vec{\omega}_2 \mid I)}, \qquad (1)$$

где T означает гипотезу, что $\vec{\omega}_1$ и $\vec{\omega}_2$ принадлежат одному диктору, а I — гипотезу, что $\vec{\omega}_1$ и $\vec{\omega}_2$ принадлежат разным дикторам. В выражении (1) значения маргинальных функций правдоподобия $P(\vec{\omega}_1,\vec{\omega}_2 \mid T)$

и $P(\vec{\omega}_i \mid I)$ могут быть посчитаны точно в случае нормального распределения скрытых переменных или приближенно — в случае распределения Стьюдента.

Так, если использовать обозначение $\mathbf{h} = (\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2)$ для всего набора скрытых переменных, то можно вычислить маргинальную функцию правдоподобия следующим образом:

$$P(\vec{\omega}) = \int P(\vec{\omega}, \mathbf{h}) d\mathbf{h} . \quad (2)$$

Поскольку в случае, если скрытые переменные имеют распределение Стьюдента, точный расчёт данного интеграла является сложной вычислительной задачей, в рамках вариационного байесовского анализа ограничиваются оценкой логарифма его нижней границы $L \le \ln P(\vec{\omega})$, используя приближенное апостериорное распределение $Q(\mathbf{h})$ для $\mathbf{h}$ скрытых параметров модели [23]:

$$L = E_h \left[\ln \frac{P(\vec{\omega}, \mathbf{h})}{Q(\mathbf{h})} \right]. \quad (3)$$

Из формулы следует, что если $Q(\mathbf{h})$ совпадает с точным апостериорным распределением $P(\mathbf{h} \mid \vec{\omega})$, тогда $L = \ln P(\vec{\omega})$, то есть оценка нижней границы (3), совпадает с логарифмом точного значения интеграла (2).

Метод на основе PLDA сравним с методом на основе JFA по значениям равновероятной ошибки EER принятия чужого и отбрасывания своего диктора, которые могут достигать величины 1–2% на фонограммах с длительностью речи более 60-90 с. Однако на практике метод PLDA, как правило, более устойчив к изменению типа канала связи и уровня шума по сравнению с методом JFA.

6. Состав системы идентификации диктора. Для участия в конкурсе было обучено 36 различных гендерно- и канало-зависимых подсистем. Подсистемы адаптировались под различные каналы получения фонограмм:

– телефонные;
– микрофонные;
– смешанные (телефон–микрофон).

Кроме того, в рамках одного канала производилось дополнительное деление подсистем на две гендерно-зависимые (для женских и мужских голосов) подсистемы.

В качестве обучающих данных были взяты речевые базы NIST SRE прошлых лет (1998-2010 гг.) общим объемом более 60 тыс. фонограмм.

Для участия в конкурсе использовалось 3 различных типа систем идентификации: на базе метода PLDA с диагональной матрицей ковариации модели UBM, на базе PLDA с полноковариационной UBM и JFA на базе диагональной ковариационной матрицы UBM. Для каждой PLDA-системы была использована нормализация, предложенная Гарсия-Ромеро в работе [24], для JFA-системы использовалась zt-нормализация [25] с использованием 300 сессий различных дикторов из базы NIST SRE 2008.

Построение мультисессионных эталонных моделей производилось следующим образом. Для метода PLDA i-вектор эталона рассчитывался как усреднение i-векторов, полученных для каждого произнесения. Для метода JFA по всем сессиям эталона рассчитывалась общая статистика Баума-Уэлша, которая затем использовалась для оценки мультисессионной модели.

Таким образом, для каждого сочетания канала связи и пола диктора было обучено 6 систем идентификации:

1) Телефонная для мужских голосов.
2) Телефонная для женских голосов.
3) Микрофонная для мужских голосов.
4) Микрофонная для женских голосов.
5) Смешанная для мужских голосов.
6) Смешанная для женских голосов.

В свою очередь, каждая система идентификации состояла из шести подсистем. Например, телефонная система представляла собой смесь следующих подсистем:

1) Подсистема PLDA (полноковариационная UBM), обученная на телефонных данных соответствующего гендера на основе MFCC признаков.
2) Подсистема PLDA (диагональная UBM), обученная на телефонных данных соответствующего гендера на основе MFCC признаков.
3) Подсистема JFA (диагональная UBM), обученная на телефонных данных соответствующего гендера на основе MFCC признаков.
4) Подсистема PLDA (полноковариационная UBM), обученная на телефонных данных соответствующего гендера на основе LFCC признаков.

5) Подсистема PLDA (диагональная UBM), обученная на телефонных данных соответствующего гендера на основе LFCC признаков.
6) Подсистема JFA (диагональная UBM), обученная на телефонных данных соответствующего гендера на основе LFCC признаков.

Микрофонная и смешанная системы также являлись смесью шести подсистем, обученных, соответственно, на микрофонных и смешанных данных.

Общее количество UBM, используемых в данной работе, было равно 24. Размерность диагональных UBM составляла 2048 смесей, полноковариационных – 1024. Количество итераций обучения UBM было равно 50.

7. Получение итогового обобщенного решения. Для повышения точности идентификации системы, голосовые модели строились различными методами, с последующим смешиванием результатов идентификации отдельными методами для получения итогового обобщенного решения.

В зависимости от типа канала связи и пола диктора на тестовой фонограмме применялась одна из 4-х обученных систем идентификации: телефонная мужская, телефонная женская, микрофонная мужская или микрофонная женская.

Результирующее решение по каждой из систем представляло собой логарифм правдоподобия (LLR) сходства диктора на спорной фонограмме с эталоном. Расчет обобщенного значения (ОР) LLR по всем подсистемам производился с помощью программы BOZARIS [26]. Схема смешивания подсистем для телефонной системы приведена на рис. 3. Телефонная система состояла из смеси 12 подсистем: 6-ти подсистем, обученных на телефонных корпусах, и 6-ти подсистем, обученных на смешанных корпусах. Применение смешанных подсистем при идентификации телефонного тестового сегмента обусловлено наличием микрофонных сессий в наборе фонограмм для обучения эталонной модели диктора.

Для обучения весовых коэффициентов ОР использовались речевые базы NIST SRE 2006-2010. Для повышения точности при смешивании подсистем в случае зашумленных тестовых сегментов, речевые базы NIST SRE 2006-2010 были искусственно зашумлены с использованием трех уровней ОСШ: без зашумления, 15 дБ и 6 дБ. ОР было обучено на каждом из уровней ОСШ. Таким образом, в зависимости от

оценки значения ОСШ тестового сегмента, применялось одна из трех настроек ОР:

1) ОСШ > 20 дБ: настройка ОР на базу без зашумления.
2) 10 < ОСШ ≤ 20 дБ: настройка ОР на базу с ОСШ 15 дБ.
3) ОСШ ≤ 10 дБ: настройка ОР на базу с ОСШ 6дБ.

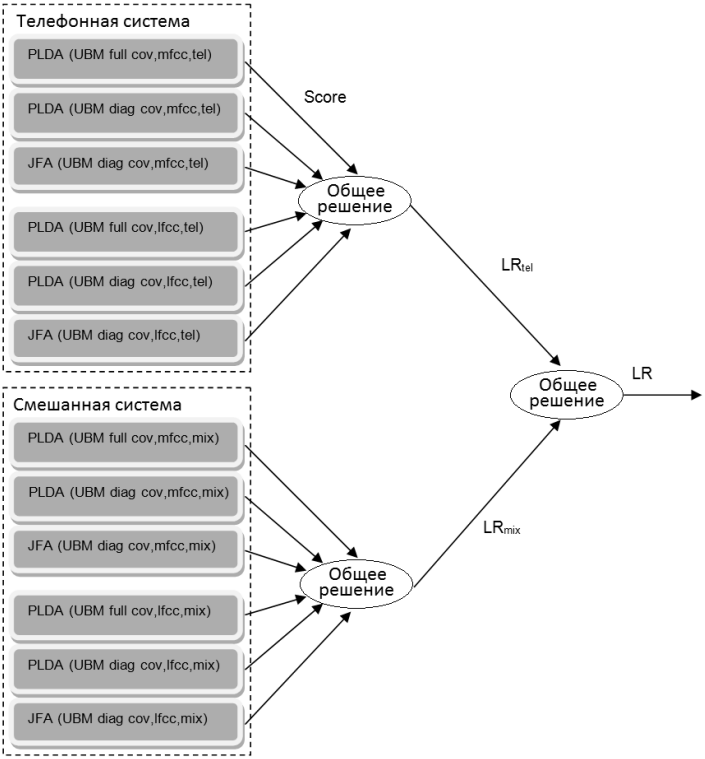

Рис. 3. Структура телефонной системы идентификации диктора ЦРТ.

7. Результаты конкурса NIST SRE 2012. Результаты конкурса NIST SRE 2012 оценивались в пяти различных номинациях, в зависимости от типа канала связи спорного произнесения (эталон во всех тестах был мультисессионным):

1) Микрофонный канал связи
2) Телефонный канал связи.
3) Микрофонный канал с добавленным шумом.
4) Телефонный канал с добавленным шумом.

5) Телефонный канал, записанный в шумных условиях.

По каждой номинации строился DET-график и оценивались два типа ошибки: minimum DCF (minDCF) – значение функции стоимости ошибки с оптимальным порогом и actual DCF (actDCF) - значение функции стоимости ошибки с порогом, установленным участником.

Значения ошибок minDCF и actDCF для каждого из тестов системы ЦРТ приведены в таблице 1.

Таблица 1. **Значения ошибок системы ЦРТ в различных номинациях**

Канал связи	minDCF	actDCF	Место среди коммерческих компаний
1. Микрофонный канал	0,354	0,444	3
2. Телефонный канал	0,291	0,541	3
3. Микрофонный канал, запись с добавлением шума	0,305	0,317	2
4. Телефонный канал, запись с добавлением шума	0,376	0,504	4
5. Телефонный канал, запись в шумных условиях	0,285	0,588	3

Сравнение результатов в микрофонном канале и микрофонном канале с зашумлением приведено на рис. 4. Не смотря на добавленный шум, во втором случае значения ошибок minDCF и actDCF оказались меньше. Это может быть объяснено различием корпусов исходных данных для проведения тестирования, а также описанной выше схемой выделения речевого сигнала. Значения ошибки EER для микрофонного канала лежит в пределах 4-5%. В данном тесте описанная система показала лучший результат среди коммерческих компаний Европы и 2-й результат в мире.

Результат тестирования системы для телефонного канала связи приведен на рис. 5. Уровень ошибки EER лежит в пределах 2,5-3,0%.

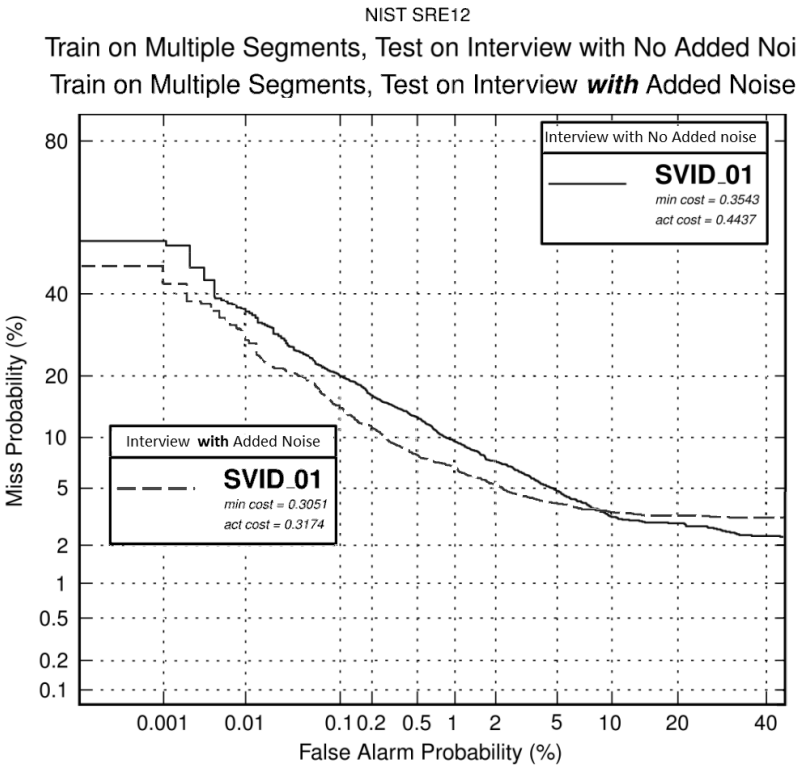

Рис. 4. DET-графики системы ЦРТ для микрофонного канала и микрофонного канала с добавленным шумом [4].

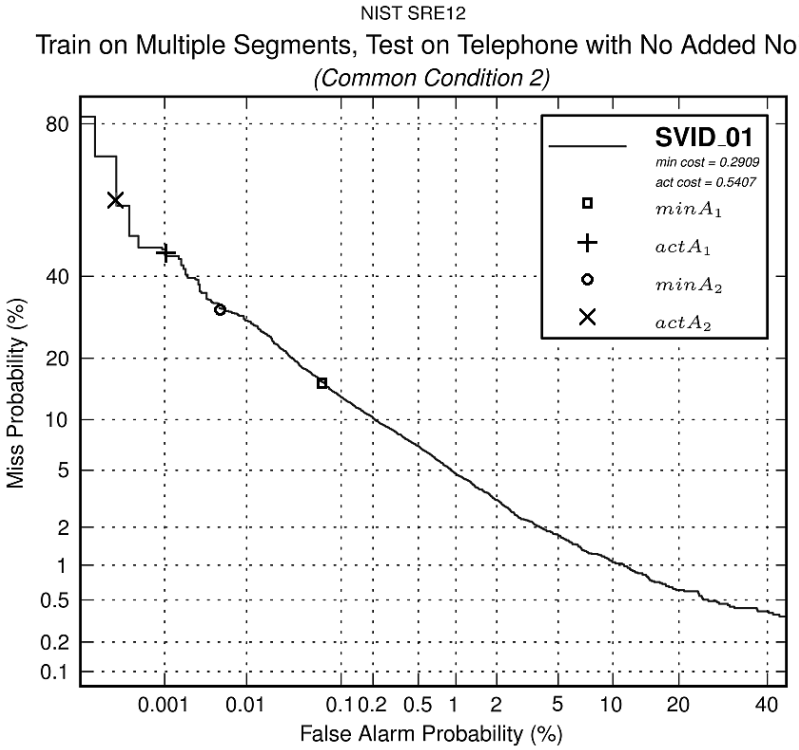

Рис.5. DET-график системы ЦРТ для телефонного канала связи [4].

8. Заключение. В рамках подготовки к конкурсу NIST SRE 2012 была произведена модернизация системы идентификации дикторов ЦРТ, использованной в конкурсе NIST SRE 2010 [1]: обучено 36 подсистем на основе методов PLDA и JFA, предложена оригинальная схема выделения речи в микрофонном канале в сочетании с собственными методами шумоочистки. В наиболее сложном тесте идентификации диктора на записях с удаленного микрофона в зашумленных условиях система ЦРТ заняла 1-ое место среди коммерческих компаний из Европы и 2-ое в мире, обогнав таких ведущих производителей как Agnitio, IBM, Cogent 3M, ValidSoft и других. В тесте идентификации диктора в телефонном канале система ЦРТ заняла 3-е место среди коммерческих компаний мира, показав уровень равновероятной ошибки EER менее 3%.

Литература

1. И.Н.Белых, А.И.Капустин, А.В.Козлов, А.И.Лоханова, Ю.Н.Матвеев, Т.С.Пеховский, К.К.Симончик, А.К. Шулипа, “Система идентификации дикторов по голосу для конкурса NIST SRE 2010”, Информ. и её примен., 6:1 (2012),С. 91–98
2. Ю.Н. Матвеев, “Технологии биометрической идентификации личности по голосу и другим модальностям”, Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение, 2012, № 3, Специальный выпуск Биометрические технологии, С. 46–61.
3. К. К. Симончик, Метод и алгоритмы текстонезависимой верификации дикторов по голосу, LAP LAMBERT Academic Publishing GmbH & Co. KG, Saarbrucken, Germany, 2011, ISBN: 978-3-8433-1295-0, 188 с.
4. “The NIST Year 2012 Speaker Recognition Evaluation Plan”, 2012, http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf.
5. “The NIST Year 2008 Speaker Recognition Evaluation Plan”, 2008, http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf.
6. “The NIST Year 2010 Speaker Recognition Evaluation Plan”, 2010, http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplan-r6.pdf.
7. К. К. Симончик, О. С. Галинина, А. И. Капустин, “Алгоритм обнаружения речевой активности на основе статистик основного тона в задаче распознавания диктора”, Научно-технические ведомости СПбГПУ, 103:4 (2010), 18–23, Издательство Политехнического университета.
8. А. В. Козлов, А. И. Лоханова, К. К. Симончик, “Алгоритм детектирования музыкальных фрагментов в задачах речевой обработки”, Научно-технические ведомости СПбГПУ, 103:4 (2010), 7–11, Издательство Политехнического университета.
9. С.В. Алейник, К.К. Симончик, “Алгоритмы выделения типовых помех и искажений в записях речевых сигналов”, Известия ВУЗов. Сер. Приборостроение, 56:2 (2013) (впечати).
10. С.В. Алейник, Ю.Н. Матвеев, А.Н. Раев, “Алгоритмы оценки уровня клиппирования речевых сигналов”, Научно-технический вестник информационных технологий, механики и оптики, 2012, №3, С. 79–83.
11. P. Ignatov, M. Stolbov, S. Aleinik, “Semi-Automated Technique for Noisy Recording Enhancement Using an Independent Reference Recording”, Audio Engineering Society Conference: 46th International Conference: Audio Forensics (Denver, CO, USA, June 14–16), 2012, С. 2–3, http://www.aes.org/e-lib/browse.cfm?elib=16342.
12. D. Reynolds, “Experimental evaluation of features for robust speaker identification”, IEEE Transactionson Speech and Audio Processing, 2:4 (1994), С. 639-643.
13. X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, S. Shamma, “Linear versus mel frequency cepstral coefficients for speaker recognition”, Proc. 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (Hawaii), 2011, С. 559–564.
14. F. Bimbot, et al., “A Tutorial on Text-Independent Speaker Verification”, EURASIP Journal on Applied Signal Processing, 2004, №4, С.430-451.
15. L. Burget, P. Matejka, O. Glembek, et al., “Analysis of feature extraction and channel compensation in GMM speaker recognition system”, IEEE Trans. on Audio, Speech and Language Processing, 15:7 (2007), С. 1979-1986.
16. P. Kenny, P. Ouellet, N. Dehak, et al., “A Study of Inter-Speaker Variability in Speaker Verification”, IEEE Transactions on Audio, Speech and Language Processing, 16:5 (2008), С. 980-988.
17. P. Kenny, G. Boulianne, P. Ouelle, P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition”, IEEE Transactions on Audio, Speech and Language Processing, 15:4 (2007), С. 1435–1447.

18. T. Pekhovsky, I. Oparin, “Eigen Channel Method for Text-Independent Russian Speaker Verification”, Speech and Computer (SpeCom’08): Proc. of the XII International Conference (Moscow), 2008, C. 385-390.
19. T. Pekhovsky, I. Oparin, “Maximum Likelihood Estimations in the Session-Independent Modelling of the Speaker”, Speech and Computer (SpeCom’09): Proc. of the XIII International Conference (St.-Petersburg), 2009, C. 267-270.
20. D.A. Reynolds, T.F. Quatieri, R.B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing, 2000, №10, C. 19-41.
21. P. Matejka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, J. Cernocky , “FullCovariance UBM and Heavy-Tailed PLDA in i-vector Speaker Verification”, Proc. ICASSP. (Prague, Czech Republic, May), 2011, C. 4828-4831.
22. K. Simonchik, T. Pekhovsky, A. Shulipa, A. Afanasyev, “Supervized Mixture of PLDA Models for Cross-Channel Speaker Verification”, 13th Annual Conference of the International Speech Communication Association, Interspeech-2012 (Portland, Oregon, USA, September 9-13), 2012.
23. V. De Silva, J.B. Tenenbaum, “Sparse multidimensional scaling using landmark points”, Technical Report (Stanford University), 2004, http://window.stanford.edu/courses/cs468-05-winter/Papers/Landmarks/Silva_landmarks5.pdf.
24. D. Garcia-Romero, C.Y. Espy-Wilson, “Analysis of I-vector Length Normalization in Speaker Recognition Systems”, Proc. of Interspeech (Florence, Italy), 2011, C.249–252.
25. R. Vogt, S. Sridharan, “Explicit Modelling of Session Variability for Speaker Verification”, Computer Speech & Language, 22:1 (2008), C.17-38.
26. “BOSARIS Toolkit”, https://sites.google.com/site/bosaristoolkit/.
27. P. Kenny,“Bayesian speaker verification with heavy tailed priors”, Keynote presentation, Proceedings of the Odyssey Speaker and Language Recognition Workshop (Brno, Czech Republic, June), 2010, http://www.crim.ca/perso/patrick.kenny/kenny_Odyssey2010.pdf.
28. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, “Front-end factor analysis for speaker verification”, IEEE Transactions on Audio, Speech & Language Processing, 19:4 (2011).

Козлов Александр Викторович — ведущий программист компании ООО «ЦРТ-инновации». Область научных интересов: идентификацией и верификацией по голосу, речевые информационные системы, биометрические системы, системы цифровой обработки сигналов и изображений, кластерные вычисления. Число научных публикаций — 3. kozlov-a@speechpro.com, www.speechpro.com; ООО «ЦРТ-инновации», ул. Красуцкого, 4, Санкт-Петербург, 196084, РФ; р.т. +7(812)325-8848, факс +7(812)327-9297.

Kozlov Alexander Viktorovich — the leading programmer, Speech Technology Center. Research area: identification and verification on a voice, speech information systems, biometric systems, systems of digital processing of signals and images, cluster calculations. The number of publications — 3. kozlov-a@speechpro.com, www.speechpro.com; ITMO, Speech Technology Center, Krasutskogo St., 4, St.Petersburg, 196084, Russia; tel. 7 (812) 325-8848, fax: +7 (812) 327-9297.

Кудашев Олег Юрьевич — программист ООО «ЦРТ-инновации», аспирант кафедры речевых информационных систем факультета Информационных технологий и программирования Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики (НИУ ИТМО). Область научных интересов: цифровая обработка сигналов, речевые технологии, идентификация дикторов по голосу, разделения дикторов по голосу. Число научных публикаций — 2.

kudashev@speechpro.com, www.speechpro.com; НИУ ИТМО, ООО «ЦРТ-инновации», ул. Красуцкого, 4, Санкт-Петербург, 196084, РФ; р.т. +7(812)325-8848, факс +7(812)327-9297. Научный руководитель — Т.С. Пеховский.

Kudashev Oleg Yuryevich — the programmer, Speech Technology Center, the graduate student of Voice Information Systems Department of Information Technology and Software St. Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO). Research area: igital processing of signals, speech technologies, identification of announcers on a voice, divisions of announcers on a voice. The number of publications — 2. kudashev@speechpro.com, www.speechpro.com; ITMO, Speech Technology Center, Krasutskogo St., 4, St.Petersburg, 196084, Russia; tel. 7 (812) 325-8848, fax: +7 (812) 327-9297. Scientific advisor — T.S. Pekhovsky.

Матвеев Юрий Николаевич — д-р техн. наук, главный научный сотрудник ООО «ЦРТ-инновации», профессор, заместитель заведующего кафедрой речевых информационных систем факультета Информационных технологий и программирования Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики (НИУ ИТМО). Область научных интересов: речевые информационные системы, мультимодальные биометрические системы, системы цифровой обработки сигналов и изображений. Число научных публикаций — 70. matveev@speechpro.com, matveev@mail.ifmo.ru, www.speechpro.com; НИУ ИТМО, ООО «ЦРТ-инновации», ул. Красуцкого, 4, Санкт-Петербург, 196084, РФ; р.т. +7(812)325-8848, факс +7(812)327-9297.

Matveev Yury Nikolaevich — the Dr.Sci.Tech., the chief researcher, Speech Technology Center, the professor, the assistant manager chair of Voice Information Systems Department of Information Technology and Software St. Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO). Research area: speech information systems, multimodal biometric systems, systems of digital processing of signals and images. The number of publications — 70. matveev@speechpro.com, matveev@mail.ifmo.ru, www.speechpro.com; ITMO, Speech Technology Center, Krasutskogo St., 4, St.Petersburg, 196084, Russia; tel. 7 (812) 325-8848, fax: +7 (812) 327-9297.

Пеховский Тимур Сахиевич — канд. физ.-мат. наук, ведущий научный сотрудник ООО «ЦРТ-инновации», доцент кафедры речевых информационных систем факультета Информационных технологий и программирования Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики (НИУ ИТМО). Область научных интересов: задачи верификации и идентификации дикторов по голосу, задача разделения дикторов в звуковом потоке, теория машинного обучения. Число научных публикаций — 20. tim@speechpro.com, www.speechpro.com; НИУ ИТМО, ООО «ЦРТ-инновации», ул. Красуцкого, 4, Санкт-Петербург, 196084, РФ; р.т. +7(812)325-8848, факс +7(812)327-9297.

Pekhovsky Timur Sakhievich — PhD, Senior Researcher, Speech Technology Center, Assistant Professor of Voice Information Systems Department of Information Technology and Software St. Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO). Research area: the task of verification and speaker identification by voice, the task division speakers in the audio stream, the theory of machine learning. The number of publications — 20. tim@speechpro.com, www.speechpro.com; ITMO, Speech Technology

Center, Krasutskogo St., 4, St.Petersburg, 196084, Russia; tel. 7 (812) 325-8848, fax: +7 (812) 327-9297.

Симончик Константин Константинович — канд. техн. наук, руководитель отдела ООО «ЦРТ-инновации», доцент кафедры речевых информационных систем факультета Информационных технологий и программирования Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики (НИУ ИТМО). Область научных интересов: цифровая обработка сигналов, верификация и идентификация дикторов по голосу, алгоритмы машинного обучения. Число научных публикаций — 18. simonchik@mail.ifmo.ru, simonchik@speechpro.com; НИУ ИТМО, ООО «ЦРТ-инновации», ул. Красуцкого, 4, Санкт-Петербург, 196084, РФ; р.т. +7(812)325-8848, факс +7(812)327-9297.

Simonchik Konstantin Konstantinovich — PhD, director of Speech Technology Center, Assistant Professor of Voice Information Systems Department of Information Technology and Software St. Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO). Research area: digital signal processing, verification and identification of speakers by voice, algorithms, machine learning. The number of publications — 18. simonchik@mail.ifmo.ru, simonchik@speechpro.com; ITMO, Speech Technology Center, Krasutskogo St., 4, St.Petersburg, 196084, Russia; tel. 7 (812) 325-8848, fax: +7 (812) 327-9297.

Шулипа Андрей Константинович — научный сотрудник компании ООО «ЦРТ-инновации». Область научных интересов: распознавание образов, программирование и математическое моделирование. Число научных публикаций — 5. shulipa@speechpro.com, www.speechpro.com; ООО «ЦРТ-инновации», ул. Красуцкого, 4, Санкт-Петербург, 196084, РФ; р.т. +7(812)325-8848, факс +7(812)327-9297.

Shulipa Andrey Konstantinovich — Researcher, Speech Technology Center. Research area: pattern recognition, programming and mathematical modeling. The number of publications — 5. shulipa@speechpro.com, www.speechpro.com; Speech Technology Center, Krasutskogo St., 4, St.Petersburg, 196084, Russia; tel. +7 (812) 325-8848, fax: +7 (812) 327-9297.

Рекомендовано лабораторией речевых и многомодальных интерфейсов, заведующий лабораторией Ронжин А.Л., д-р техн. наук, доцент.
Статья поступила в редакцию 05.02.2013.

РЕФЕРАТ

Козлов А.В., Кудашев О.Ю., Матвеев Ю.Н., Пеховский Т.С., Симончик К.К., Шулипа А.К. **Система идентификации дикторов по голосу для конкурса NIST SRE 2012.**

В статье представляется описание текстонезависимой системы автоматической идентификации дикторов по голосу, разработанной ООО «ЦРТ» для участия в международном конкурсе по оцениванию систем распознавания дикторов NIST SRE 2012. Начиная с 1996 года, этот конкурс каждые два года проводится Национальным институтом стандартов и технологий США. Его основная цель – оценить уровень существующих технологий распознавания дикторов и определить перспективные направления развития индустрии.

Представленная ООО «ЦРТ» система текстонезависимой идентификации диктора основана на выделении речи из фонограмм и последующем попарном сравнении биометрических признаков (содержащихся в голосе индивидуальных, идентификационно значимых признаков личности). Для выделения речевых сегментов использовались специальные алгоритмы предобработки всего сигнала. В частности, для выделения речи в микрофонном канале предложена оригинальная схема выделения речи целевого диктора на основе стереофильтрации. В качестве речевых признаков использовались векторы мел-частотных и линейных кепстральных коэффициентов. В качестве методов идентификации использовались наиболее распространенные подходы на основе использования моделей гауссовых смесей: JFA и PLDA. В работе сделан упор на описание математических основ и преимуществ данных подходов.

Результаты конкурса NIST SRE 2012 оценивались в пяти различных номинациях, в зависимости от типа канала связи: микрофонный, телефонный, микрофонный с добавленным шумом, телефонный с добавленным шумом и телефонный канал, записанный в шумных условиях. Для повышения точности идентификации системы, голосовые модели строились различными подсистемами, с последующим смешиванием результатов идентификации для получения итогового обобщенного решения. В зависимости от типа канала связи и пола диктора на тестовой фонограмме применялась одна из 4-х обученных систем идентификации: телефонная мужская, телефонная женская, микрофонная мужская или микрофонная женская, каждая из которых состояла из 12 подсистем на основе методов PLDA и JFA.

В наиболее сложном тесте идентификации диктора на записях с удаленного микрофона в зашумленных условиях система ЦРТ заняла 1-ое место среди коммерческих компаний из Европы и 2-ое в мире, обогнав таких ведущих производителей как Agnitio, IBM, Cogent 3M, ValidSoft и других. В тесте идентификации диктора в телефонном канале система ЦРТ заняла 3-е место среди коммерческих компаний мира, показав уровень равновероятной ошибки EER менее 3%.

SUMMARY

Kozlov A.V., Kudashev O.Yu., Matveev Yu.N., Pekhovsky T.S., Simonchik K.K., Shulipa A.K. **Speaker recognition system for the NIST SRE 2012.**

In the paper the description of the text-independent system of automatic speaker recognition on base of their voice, developed by STC Ltd for the international competition on evaluation of speaker recognition systems NIST SRE 2012 is submitted. Since 1996, this competition is held each two years by National institute of standards and technologies of the USA. Its main objective – to estimate the level of existing technologies of speaker recognition and to define the perspective directions of development of the industry.

The STC's system of the text-independent speaker recognition is based on speech extraction from phonograms and the subsequent pairwise comparison of biometric features (the identification of the significant personality traits contained in the voice of the individual). To separate speech segments special algorithms of preprocessing of whole signal were used. In particular, to separate speech in the microphone channel the original scheme of speech extraction of the target speaker on the basis of a stereo filtration is offered. As speech features vectors were used mel-frequency cepstral coefficients and linear-frequency cepstral coefficients. As methods of identification the most widespread approaches on the basis of use of the Gaussian mixed models were used: JFA and PLDA. The paper focuses on the description of mathematical principles and benefits of these approaches.

Results of the competition 2012 NIST SRE were evaluated in five different categories, depending on the type of communication channel: microphone, telephone, microphone with added noise, telephone with the added noise and telephone in noisy environments. To improve the accuracy of the identification system, voice models were built by different subsystems, with the subsequent fusion of results of identification for obtaining the generalized decision. Depending on the type of channel and the gender of the speaker on the test recording one of 4 trained identification systems was applied: telephone male, telephone female, microphone male or microphone female, each of which consisted of 12 subsystems based on the PLDA and JFA methods.

In the most difficult test of speaker recognition on recordings from a remote microphone in noisy conditions, the STC's system took first place among commercial companies from Europe and the second in the world, ahead of such leading manufacturers as Agnitio, IBM, Cogent 3M, ValidSoft and others. In the test of speaker identification in telephone channel STC's system took third place among the commercial companies of the world, showing an equal error rate less than 3%.