УДК 004.93+57.087.1

Ю.Н. Матвеев

ТЕХНОЛОГИИ БИОМЕТРИЧЕСКОЙ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ ПО ГОЛОСУ И ДРУГИМ МОДАЛЬНОСТЯМ

Рассмотрены достижения в области автоматических методов идентификации личностей по голосу, которые позволили приблизить рабочие характеристики голосовой модальности к характеристикам других биометрических модальностей, в особенности к лицевой. Приведен метод мультиалгоритмического и мультимодального смешивания на уровне оценок при совместном использовании нескольких биометрических характеристик различной модальности. Приведены экспериментальные данные построения обобщенного решения по нескольким модальностям.

E-mail: matveev@speechpro.ru

Ключевые слова: *идентификация личности, голосовая модальность, мультимодальная биометрическая система.*

Введение. В связи с бурным развитием биометрических технологий в мире, особое место в исследованиях ООО «Центр речевых технологий» (ЦРТ) занимает разработка инновационных решений по голосовой биометрии. Решения, полученные в ЦРТ, для создания и ведения фоноучетов, проведения автоматической идентификации личности по голосу основаны на таких методах автоматического исследования голоса и речи, для которых не имеет значения язык речи, акцент диктора и используемый диалект, а также содержание текста произносимой речи.

Уникальность голосовой биометрии состоит в том, что это единственная биометрическая модальность, которая позволяет идентифицировать человека по телефону. Это важно, например, при удаленном доступе к различным услугам, при криминалистической идентификации, где единственным доказательством является запись телефонного разговора подозреваемого. Кроме того, голосовая идентификация не требует применения специализированного дорогостоящего оборудования. Все, что необходимо – обычный микрофон. При этом по уровню надежности голосовая биометрия не уступает, а по некоторым характеристикам превосходит характеристики других систем биометрической идентификации.

Следует отметить, что недавно системы биометрической идентификации по голосу обладали значительно худшими рабочими характеристиками (точность идентификации, размер биометрической модели и т. д.), по сравнению с рабочими характеристиками систем

биометрической идентификации других модальностей. Однако за последние три-четыре года в области разработки автоматических методов голосовой идентификации были достигнуты значимые успехи, которые позволили приблизить рабочие характеристики голосовой модальности к рабочим характеристикам других модальностей, в особенности к лицевой (табл. 1) [1]. Это привело к тому, что голосовая модальность теперь не является «тормозом» при создании мультимодальных систем биометрической идентификации личности.

Таблица 1

Значения ошибок идентификации для различных биометрических модальностей

Биометрический признак	Тест	Условия тестирования	FRR, %	FAR, %
Отпечатки пальцев	FVC 2006	Неоднородная популяция (включает работников ручного труда и пожилых людей)	2,2	2,2
Лицо	MBE 2010	Полицейская база фотографий База фотографий с документов	4,0 0,3	0,1 0,1
Голос*	NIST 2010	Текстонезависимое распознавание	3…4	1,0
Радужная оболочка глаз	ICE 2006	Контролируемое освещение, широкий диапазон качества изображений	1,1…1,4	0,1
* Результаты получены в ЦРТ.				

Размеры моделей, кбайт, для различных биометрических модальностей приведены ниже (в ЦРТ достигнут размер модели 2…3 кбайт):

Голос ……. ..... 70…80
Лицо ……... ..... 0,1…2,0
Подпись …..... 0,5…1,0
Отпечатки пальцев …..... 0,25…1,20
Геометрия руки ……. 0,01
Радужная оболочка глаз …. ..... 0,25…0,50
Сетчатка глаз ……. 0,1

Дополнительно к голосовой модальности выбирается лицевая модальность, что объясняется широким распространением соответствующих бимодальных устройств (сотовые телефоны, коммуникаторы, цифровые фотокамеры и видеокамеры, ноутбуки). Наличие таких бимодальных устройств упрощает процесс получения биометрических образцов, процесс регистрации личности в системе биометрической идентификации, снижает стоимость самой системы и т. д.

Для совместного использования биометрических характеристик одной из различных модальностей были исследованы и разработаны методы мультиалгоритмического и мультимодального смешивания (построения обобщенного решения по нескольким признакам одной или нескольких модальностей).

Особенности биометрической идентификации личности по голосу. Уникальность голоса человека обусловлена множеством физиологических особенностей (строением голосовых связок, трахеи, носовых полостей, манерой произношения звуков, расположением зубов). Комбинация этих особенностей индивидуальна, как и отпечатки пальцев. Однако на практике ни одна из унимодальных систем биометрической идентификации, в том числе и голосовая, не может гарантировать 100 %-ной идентификации личности.

Основными источниками ошибок при идентификации дикторов являются эффекты:

– среды записи (уровень и тип шума, уровень реверберации);

– представления (длительность речи, психофизиологическое состояние говорящего (болезнь, эмоциональное состояние и т. п.), язык речевого сообщения, изменение голосового усилия);

– канала (помехи (импульсные, тональные и т. п.), искажения (амплитудно-частотные характеристики микрофона и канала передачи, вид кодирования в канале и т. д.)).

Для снижения влияния перечисленных источников ошибок при проектировании системы голосовой биометрии в ЦРТ были созданы робастные автоматические методы и алгоритмы, реализующие основные этапы обработки речевого сигнала в такой системе:

– предварительная обработка речевого сигнала (выделение на речевом сигнале участков, содержащих речь дикторов, оценка качества речевого материала);

– автоматическая сегментация дикторов в фонограмме;

– автоматическое выделение биометрических характеристик голоса и речи;

– идентификация дикторов.

Успехи ЦРТ в области голосовой биометрии были достигнуты благодаря тщательной научной и технической проработке каждого из указанных выше этапов.

Предварительная обработка речевого сигнала. На рис. 1 приведена схема предварительной обработки речевого сигнала, используемая в биометрических решениях ЦРТ. В любой системе обработки речи необходимо, прежде всего, выделить из входного сигнала речевые фрагменты, отбросив паузы и участки, содержащие различные виды помех. В контексте данной задачи помехами, которые необходимо детектировать и исключить из дальнейшего анализа, могут быть щелчки, гудки, DTMF-сигналы, музыкальные фрагменты, ха-

рактерные для телефонных каналов связи. В ЦРТ разработан алгоритм детектирования музыкальных фрагментов на основании динамики спектральных максимумов [2]. Экспериментальные исследования представительной базы показали, что этот алгоритм имеет бóльшую эффективность, чем существующие аналоги. Малое значение ошибки (11 %), а также простота данного алгоритма удовлетворяют требованиям реальных приложений.

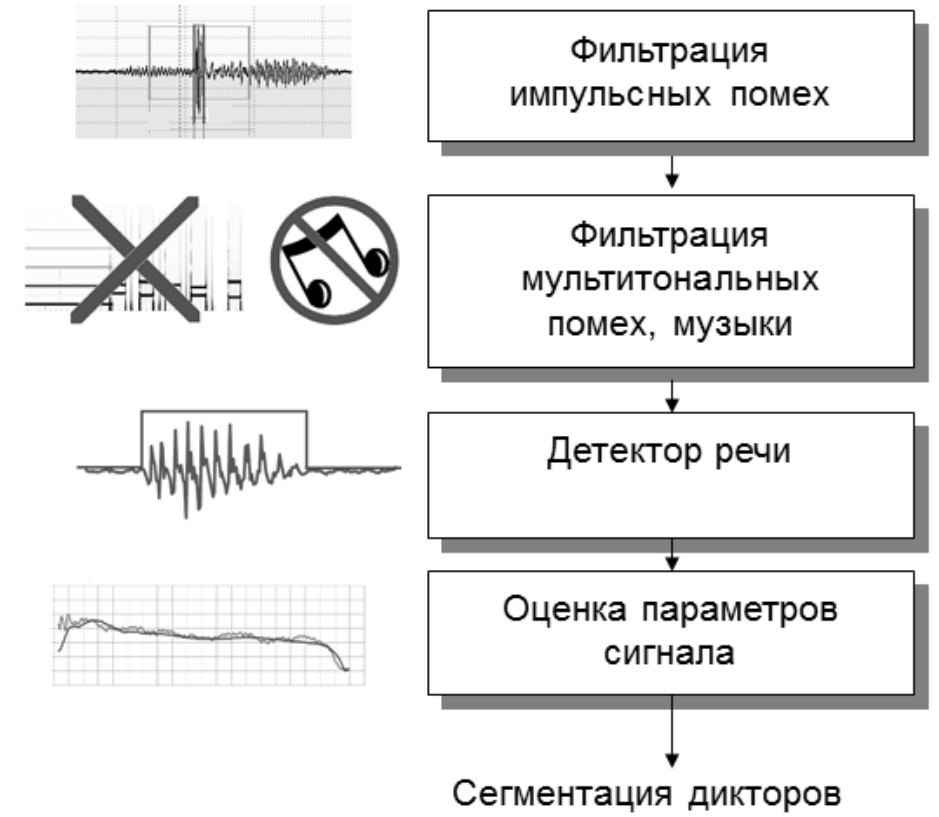

Рис. 1. Схема предварительной обработки речевого сигнала (выделение на речевом сигнале участков, содержащих речь дикторов)

Важным компонентом является детектор речевой активности (Voice Activity Detector, VAD). Основное внимание при разработке VAD-алгоритма уделяется выделению шумоустойчивых признаков и выбору правил классификации речь – не речь. Как правило, используются алгоритмы на основе анализа энергии сигнала, обнаружения основного тона, спектрального и кепстрального анализа, измерений числа переходов сигнала через нуль [3, 4]. Несмотря на значительное количество реализаций VAD-алгоритма и многочисленные исследования, существующие решения не полностью отвечают требованиям, продиктованным особенностями задачи идентификации личности по голосу.

В ЦРТ создан VAD-алгоритм (модификация алгоритма на основе статистик основного тона [3]), который выделяет вокализованные участки речи [5]. Главная идея выделения указанных участков речи заключается в использовании гласных и назализованных согласных. С одной стороны, недостатком является потеря некоторых согласных, с другой, – взрывные согласные и аффрикаты обладают меньшей идентификационной значимостью. Тогда можно предположить, что потеря некоторой части незначимого речевого материала будет компенсироваться качественным удалением неречевых участков. Это позволяет, например, снизить зависимость качества идентификации диктора от искажений канала в паузах. В основе разработанного VAD-алгоритма лежит спектральный анализ речевого сигнала. На каждом кадре спектрограммы осуществляется поиск положений максимумов, соответствующих гармоникам основного тона, по которым оценивается значение его частоты. При этом в сигнале возможно отсутствие нижних гармоник основного тона, что характерно для телефонного канала с полосой частот 300…3400 Гц. Работа VAD-алгоритма проверяется наложением полученных кривых основного тона на спектрограмму (рис. 2). Таким образом, можно выявить следующие преимущества использования VAD-алгоритма на основе анализа частоты основного тона: выделение речевого сигнала происходит, в том числе на относительно зашумленных участках (соотношение сигнал – шум 10 дБ и ниже); непрерывность значения основного тона и принадлежность этого значения области диапазону значений частоты, типичных для речи.

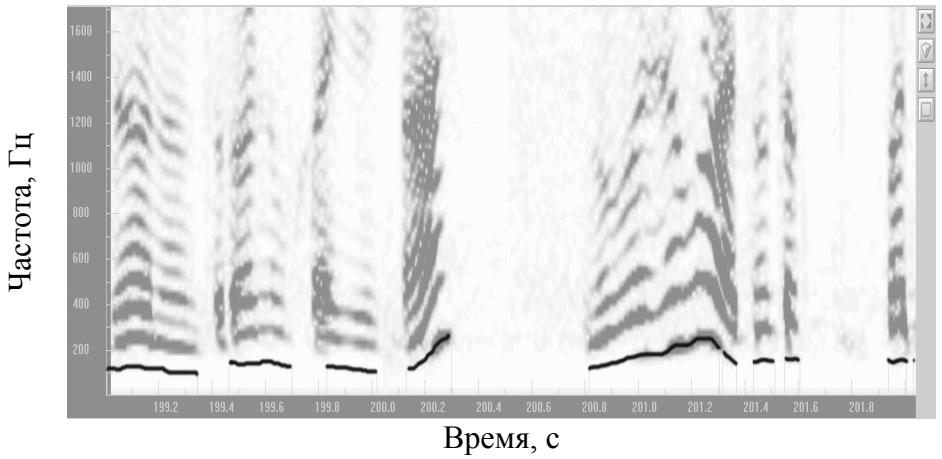

Рис. 2. Наложение кривых основного тона на спектрограмму

Экспериментально установлено, что VAD-алгоритм обеспечивает снижение EER системы идентификации на кепстральных признаках в 2 раза (с 12 % для энергетического детектора до 6 % для VAD-алгоритма). При использовании в качестве информационных акусти-

ческих признаков мел-кепстральных коэффициентов (Mel Frequency Cepstral Coefficients, MFCC) система идентификации чувствительна, в первую очередь, к ошибке ложного детектирования шума как речи. Применение VAD-алгоритма предъявляет более жесткие требования к частотной характеристике сигнала, исключая помехи большой амплитуды, не являющиеся речью.

На последнем шаге этапа предварительной обработки оценивается качество речевого сигнала на выделенных участках и принимается решение о возможности идентификации личности по данному речевому материалу или об отказе от идентификации.

Автоматическая сегментация дикторов в фонограмме. На фонограммах, записанных в реальных условиях, типовыми являются следующие случаи (рис. 3): наложение различных акустических помех (от телевизора, радио и т. п.) на речь дикторов; наличие на фонограмме речи нескольких дикторов; наложение речи нескольких дикторов друг на друга и образование так называемого голосового коктейля.

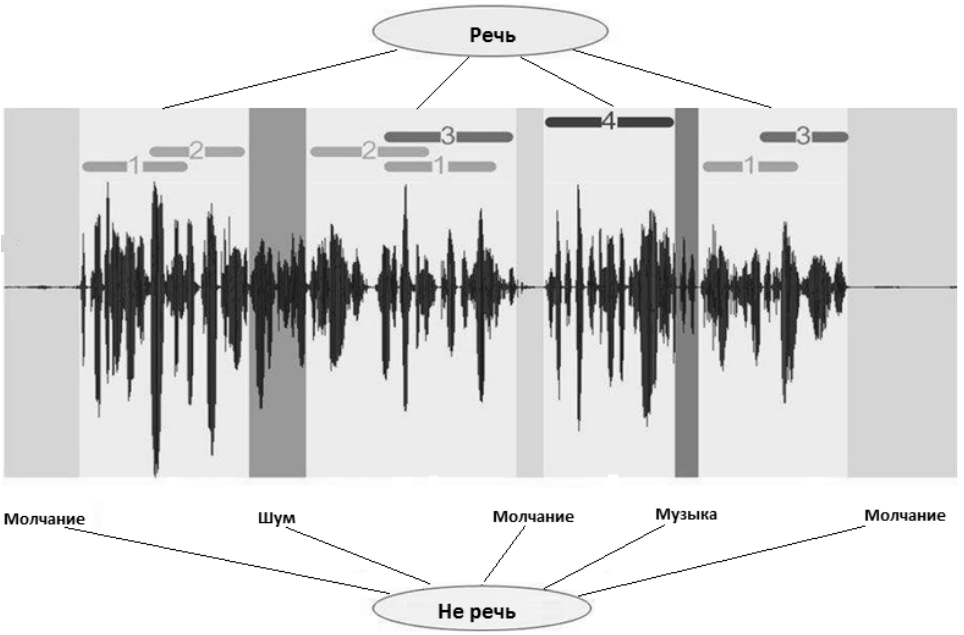

Рис. 3. Схема предварительной обработки речевого сигнала (сегментация дикторов):

1–4 – номера дикторов, речь которых содержится в фонограмме

Для решения перечисленных случаев сегментации в ЦРТ созданы технологии:

– выделения в фонограмме речи диктора на фоне акустических помех, где для подавления помехи и выделения речи используется

образец соответствующей помехи, взятый из Интернета, компакт-диска и т. д.;

– разделения речи дикторов в голосовом коктейле по частоте основного тона;

– разметки выделенных участков речевого сигнала по принадлежности различным дикторам (определение кто и когда говорит), так называемая диаризация речи дикторов.

Задача диаризации речи дикторов имеет ряд ограничений: в большинстве случаев число дикторов в фонограмме неизвестно; отсутствуют голосовые модели дикторов; различаются объем речи, порядок и частота смены дикторов. В настоящее время были достигнуты показатели надежности (Diarization Error Rate, DER):

– известных коммерческих решений – 8…12 %;

– решений ЦРТ – 5…6 %;

– наилучших достижений – 2…3 %.

Достигнутая скорость вычислений решений ЦРТ составляет не менее 30 RT (CPU Intel i5, частота 2,8 ГГц, одноядерный процессор), что в 2–3 раза превышает значения известных скоростей. Результаты получены за счет использования гибридной дискриминационно-порождающей EV–HMM(GMMs)-системы диаризации, которая превзошла по качеству диаризации наилучшую на текущий момент EV–VBA-систему в случае коротких диалогов [6, 7].

Автоматическое выделение биометрических признаков голоса и речи. В биометрических системах ЦРТ внедрены методы автоматического выделения традиционно используемых экспертами акустических признаков: частоты основного тона диктора (частоты смыкания – размыкания голосовых связок) и формантные частоты (резонансные частоты голосового тракта) [8, 9]. Для использования в статистических методах идентификации реализовано автоматическое выделение различных кепстральных признаков: MFCC, линейных по частоте кепстральных коэффициентов (Linear Frequency Cepstral Coefficients, LFCC), кепстральных коэффициентов линейного предсказания (Linear Prediction Cepstral Coefficients, LPCC) и т. д. [10].

В статистических методах идентификации модель голоса диктора представляет собой аппроксимацию распределения используемых признаков смесью гауссовых распределений (GMM-модель).

Рассмотрим построение GMM-модели [11]. Для D-мерного вектора признаков x функция плотности распределения имеет вид

$$p(x \mid \lambda) = \sum_{i=1}^{M} \omega_i p_i(x),$$

где M – количество компонент смеси; ω_i – вес i-й компоненты смеси; $p_i(x)$ – плотность распределения i-й компоненты смеси.

Плотность распределения *i*-й компоненты смеси представляет собой *D*-мерный Гауссиан:

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

где μ – вектор математического ожидания; σ – ковариационная матрица.

Плотность распределения смеси гауссовых распределений

$$\lambda = \{\omega_i, \mu_i, \sigma_i\}.$$

Идентификация дикторов. Процедура поиска интересующего диктора (идентификации личности по голосу) заключается в автоматическом попарном сравнении голосовых моделей, в которых закодированы индивидуальные (биометрические) характеристики голоса и речи дикторов. По результатам сравнения выводится ранжированный список фонограмм, содержащих с указанной вероятностью речь интересующих дикторов.

В системах ЦРТ поиск может осуществляться с помощью метода анализа статистик основного тона (PS), спектрально-формантного (SF) метода, GMM–SVM-метода [3, 9, 10]. Наиболее распространенным является GMM–SVM-метод (рис. 4).

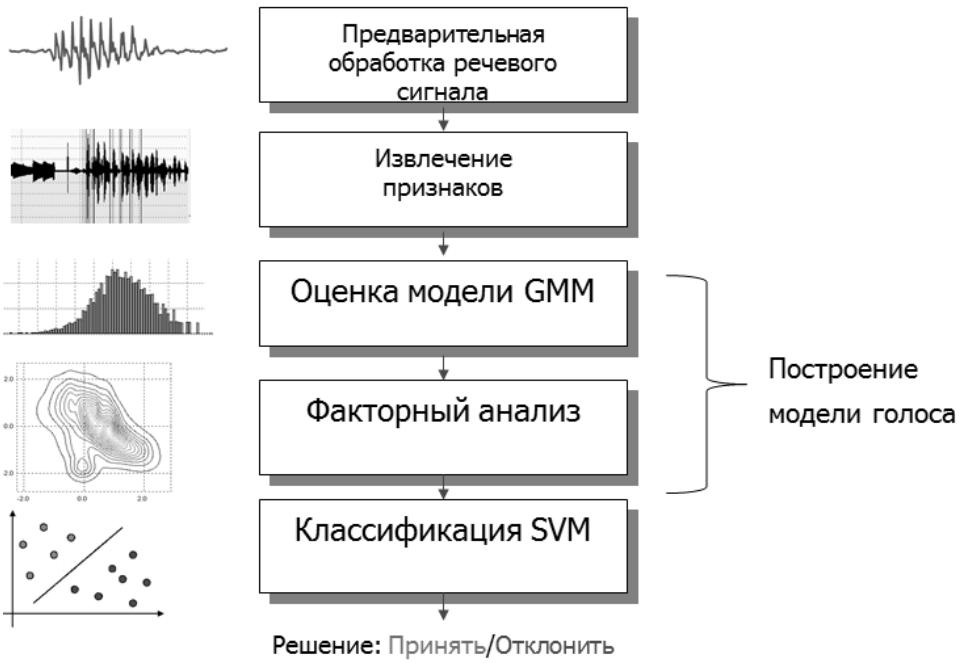

Рис. 4. Схема обработки речевого сигнала GMM–SVM-методом

Главная проблема при решении задачи текстонезависимого распознавания диктора – рассогласование, вызванное изменчивостью сессий записи голоса отдельного диктора. Причинами этого рассогласования могут быть шумы окружающей среды при записи, искажения в каналах записи и передачи речевого сигнала, а также изменчивость голоса самого диктора. Учет эффектов канала – самый значимый фактор из перечисленных выше. Для решения указанной проблемы традиционным стало применение совместного факторного анализа (Joint Factor Analysis, JFA), который позволяет эффективно расщеплять дикторскую и канальную информацию в отдельном произнесении диктора [12]. В свою очередь, это дает возможность строить каналонезависимые GMM-модели речи диктора и подавлять эффекты канала в тестовом произнесении.

Модели голоса в JFA имеет вид

$$M = m + Ux + Vy + Dz,$$

где M – супервектор GMM-модели фонограммы; m – супервектор универсальной фоновой модели (Universal Background Model, UBM); U, V, D – матрицы собственных каналов (Eigen Channel), собственных голосов (Eigen Voice), остаточной изменчивости соответственно; x, y, z – скрытые векторы.

Следует отметить, что для построения полной JFA-модели требуется большой объем обучающей базы данных. Для оценки влияния объема такой базы на надежность идентификации дикторов по голосу были проведены эксперименты (табл. 2). Система идентификации обучалась на признаках MFCC 39 (13 + 13Δ′ + 13Δ′′, где Δ′, Δ′′ – первая и вторая производная признаков MFCC), без учета эффектов канала, на мужском корпусе речевых данных. Из результатов ясно, что ошибка идентификации уменьшается с увеличением объема обучающей базы данных.

Таблица 2

Значения надежности идентификации диктора по голосу в зависимости от объема обучающей базы данных

Объем обучающей базы данных (количество файлов)	Число компонент GMM-модели	Размер матрицы собственных голосов	EER, %
515	256	30	15,6
1110	512	100	7,8
17 000	2000	300	5,8

Кроме того, в JFA обязательно используется UBM-модель, цель построения которой охарактеризовать «чужих» дикторов во всех возможных контекстах. Обучающая база UBM формируется с учетом максимально большого объема речевых данных, сбалансированных по гендерному типу, каналам записи, акустическим условиям и т. д. Как правило, в настоящее время применяется стандартная процедура построения UBM-модели, основанная на оценке максимального правдоподобия (Maximum Likelihood, ML) – ML-метод [13].

Задача ML-метода – нахождение по заданному числу T обучающих векторов данных $X = \{x_1, x_2, \ldots, x_T\}$ параметров модели λ, максимизирующих функцию правдоподобия модели:

$$p(X \mid \lambda) = \sum_{t=1}^{T} p(x_t \mid \lambda).$$

Поскольку функции параметров модели λ не линейны и напрямую максимизировать функцию правдоподобия невозможно, то используются приближенные значения оценок максимального правдоподобия, полученные с помощью EM-алгоритма (Expectation–Maximization). Существуют различные варианты этого алгоритма, где одновременно обучается набор из 512, 1024 или 2048 гауссовых компонент, и заканчивая более сложной процедурой с последовательным расщеплением компонент в процессе обучения.

С увеличением объемов речевых баз данных, наиболее важной проблемой при построении UBM-модели является поиск точного соответствия числа компонент UBM-модели количеству обучающего материала. При обучении по оценке максимального правдоподобия делается попытка определить параметры всех гауссоид. При небольшом количестве обучающего материала происходит эффект переобучения GMM-модели и снижение эффективности системы идентификации дикторов по независимой тестовой выборке, демонстрируется ее плохая обобщающая способность. При большом количестве обучающего материала выбранное число компонент UBM-модели может быть меньше оптимального их числа, разрешенного объемом обучающей базы данных, при котором система идентификации могла бы показать лучшую эффективность. Возникает вопрос определения оптимального количества компонент UBM-модели для имеющегося обучающего речевого материала.

Стандартный ответ на вопрос – использование кроссвалидационного подхода, в котором сначала проводится обучение последовательного ряда UBM-моделей с различным числом компонент, а затем – тестирование по независимой выборке систем верификации с применением каждой из UBM-моделей [14]. Такой подход в вычислительном отношении очень сложен и длителен по времени.

В ЦРТ было предложено при построении UBM-моделей использовать вариационный байесовский анализ (Variational Bayesian Analysis, VBA), который имеет следующие преимущества [15]:

– высокая устойчивость системы идентификации дикторов при переходе на другой тестовый материал за счет решения проблемы переобучения;

– высокая эффективность системы идентификации дикторов путем предотвращения выбора заведомо заниженных размеров моделей для используемого обучающего материала.

Результаты тестирования VBA на системе, представляющей собой классическую GMM-модель (Baseline-GMM), где модели речи диктора получаются путем адаптации UBM-модели по принципу максимума апостериорной вероятности (Maximum a Posteriori, MAP) приведены ниже [11]. Для оценки меры близости моделей использовалось отношение правдоподобия. В результате VBA-обучения UBM-модели сохранено 708 компонент для малого объема обучающей базы данных и 3062 компоненты для большого объема обучающей базы данных.

Значения EER, %, при различных вариантах обучения UBM-модели

ML-обучение UBM-модели ……	6,38 (M = 1024)	15,55 (M = 2048)
VBA-обучение UBM-модели ……	5,89 (M = 708)	14,24 (M = 3062)

Для малого объема базы данных при ML-обучении UBM-модели применяется большее число гауссовых распределений, чем это необходимо для описания обучающей выборки. Завышенное число гауссовых распределений приводит к переобучению UBM-модели на данных, что ухудшает результаты тестирования на другой базе. Для большого объема обучающей базы, сопоставимого с объемом обучающего материала современных систем идентификации диктора, оказалось, что обычно используемое число гауссовых распределений (2048) заведомо занижено. Тем самым, потенциал систем идентификации снижается на 8…10 % относительно результатов с VBA-обучением UBM-модели.

Кроме порождающего GMM-метода, в системах ЦРТ используется дискриминантный метод распознавания речи диктора – машины опорных векторов (Support Vector Machine, SVM). Исследования ЦРТ показали, что гибридная SVM–GMM-система обладает лучшей эффективностью, чем отдельно взятые системы, как по параметрам точности, так и по параметрам быстродействия. Применение SVM-метода делает гибридную GMM–SVM-систему более робастной к различным шумам, а также к межсессионной и внутридикторской вариативности.

Разработанный в ЦРТ вариант гибридной GMM–JFA–SVM-системы, где SVM-метод используется не в пространстве акустических векторов, а в модельном пространстве супервекторов средних GMM-моделей, является одним из самых эффективных методов в настоящее время. По данным Национального института стандартизации США, эта система вошла в число лидеров среди систем идентификации дикторов NIST SRE 2010 [1]. Следует отметить, что в такой системе для SVM-метода использовались различные линейные и нелинейные ядра, а также их линейные комбинации.

Несмотря на преимущества GMM–JFA–SVM-системы, она обладает одним недостатком – большим объемом модели голоса (50…100 кбайт), что неприемлемо при построении крупномасштабных систем биометрической идентификации.

Решить эту проблему можно с помощью низкоразмерных векторов признаков. Так, в одной из версий JFA для генерации векторов признаков используется матрица полной изменчивости (Total Variability, TV) – TV-метод [16].

Модель голоса в TV-методе имеет вид

$$M = m + Tw,$$

где w – низкоразмерный вектор; T – матрица полной изменчивости.

TV-метод – наиболее перспективный метод идентификации дикторов, обеспечивающий изящный способ редуцирования высокоразмерных входных данных к низкоразмерному вектору признаков, сохраняя большую часть полезной информации. Это снижает объем модели голоса диктора до 2…3 кбайт, что уже приемлемо при построении крупномасштабных систем биометрической идентификации. Отметим также, что GMM–TV–SVM-система позволяет получить хорошие характеристики даже для коротких произнесений.

Мультиалгоритмическое и мультимодальное смешивание. Для повышения надежности биометрической идентификации дикторов в ряде систем ЦРТ, вместе с мультимодальным смешиванием, реализовано мультиалгоритмическое смешивание с принятием обобщенного решения по нескольким или всем методам идентификации личности по голосу. Учитывая разное поведение методов в условиях, зависящих от типов каналов, длительности речевых сигналов, уровней шума и реверберации в речевых сигналах, при мультиалгоритмическом смешивании получаем значительно более высокую и стабильную надежность.

Построение обобщенного решения реализуется по схеме, приведенной на рис. 5. Далее описан алгоритм получения обобщенного решения бимодальной системы ЦРТ. Рассматриваются гипотезы:

H_0: сравнение биометрических характеристик одного человека;

H_1: сравнение биометрических характеристик разных людей.

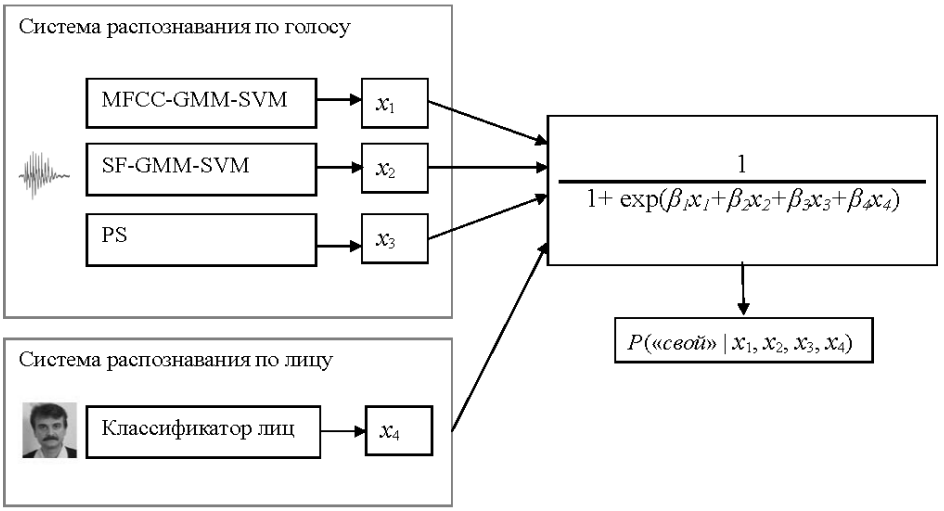

Рис. 5. Схема построения обобщенного решения

Требуется определить критерий для проверки гипотез с использованием решения классификаторов по голосовой и лицевой модальностям. Для этого объединяются выходы классификаторов $x = (x_1, x_2, x_3, x_4)$ для получения более объективного решения.

Обобщенное решение заключается в том, что сравнение принадлежит к классу свой – свой с вероятность $P(H_0|x)$.

Апостериорные вероятности $p(H_i|x_j)$ для каждого j-го классификатора оцениваются на тестовой базе по описанной далее схеме.

Решение каждого классификатора в отдельности представляет собой значение классифицирующей функции. Таким образом, решается задача вероятностного оценивания: наряду с классификацией объекта $a(x)$ вычисляются оценки вероятностей $p(H_i|x_j)$ для всех классов. Основным методом решения для взвешенного голосования выбрана калибровка выходов классификатора.

Для оценки вероятностей $p(H_i|x_j)$ используется аппроксимация диаграммы надежности сигмоидальным преобразованием, основанная на оценке максимального правдоподобия:

$$p(H_i \mid x_j) = \frac{1}{1+\mathrm{e}^{A_j x_j + B_j}}, \qquad (1)$$

где A_j и B_j – параметры, подбираемые по тестовой выборке; x_j – выход j-го классификатора.

Предполагается, что на этапе обучения классификатора известна конечная совокупность прецедентов, пар объект – ответ, называемая обучающей выборкой. На основе этих данных требуется восстано-

вить зависимость, т. е. построить алгоритм, способный классифицировать произвольный объект. Для этого формируется алгоритм классификации, обеспечивающий минимальное значение функционала среднего риска

$$r = P(H_0)\mathrm{FRR} + P(H_1)\mathrm{FAR}.$$

Мера близости по обобщенному решению представляет собой следующую взвешенную комбинацию:

$$S = \sum_{j=1:4} w_j \ln \frac{p(H_0 \mid x_j)}{p(H_1 \mid x_j)}, \tag{2}$$

где w_j – коэффициенты, отражающие степень доверия для каждого метода идентификации,

$$w_j = \frac{1}{2} \ln \frac{1 - Q_j}{Q_j}.$$

За оценку качества работы метода идентификации была принята его равновероятная ошибка $Q_j = \mathrm{EER}_j$. Подставляя в (2) выражение (1) для каждой вероятности $P(H_i \mid x_j)$, получаем: $S(x) = \alpha_1 x_1 + {}$ $+ \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha$, где α_i – коэффициенты, выражаемые через веса ω_j методов и параметры A_j и B_j каждого метода.

На последнем этапе вычисляется апостериорная вероятность гипотезы H_0 по результатам идентификации всеми методами. Для этого осуществляется калибровка обобщенного решения

$$P(H_0 \mid x) = \frac{1}{1 + \exp\left[AS(x) + B\right]} = \\ = \frac{1}{1 + \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_0)}. \tag{3}$$

Выражение (3) представляет собой формулу логистической регрессии.

Для проверки разработанного метода смешивания провели эксперименты на базе данных биометрических образцов (50 мужчин и 50 женщин). В качестве эталонов выбирались фонограммы из базы «ЦРТ-Микрофон» и изображения из базы ЦРТ лиц (фото с документов). За спорные фонограммы принимались фонограммы из баз «ЦРТ-Микрофон» и ЦРТ–GSM, а за спорные изображения – изображения из базы ЦРТ лиц (фото с пропуска). Результаты экспери-

ментов, приведенные ниже, подтверждают эффективность метода смешивания различных модальностей:

	EER, %	FRR, %, при FRA=10 %
Лицо ……...	5,9	3,2
Голос …….....	2,2	0,3
Обобщенное решение ……...	0,1	0

Заключение. В последнее время были достигнуты значительные успехи в идентификации личности по голосу и другим модальностям. Однако исчерпаны далеко не все резервы по повышению надежности биометрической идентификации личности. Так, перспективными направлениями развития идентификации личности являются повышение качества предварительных исходных биометрических образцов; извлечение более робастных идентификационных признаков и их комбинаций; реализация мультимодального смешивания не на уровне оценок, а на уровне признаков различной модальности.

СПИСОК ЛИТЕРАТУРЫ

1. Матвеев Ю. Н., Симончик К. К. Система идентификации дикторов по голосу для конкурса NIST SRE 2010 // ГрафиКон'2010. Тр. 20-й Межд. конф. по компьютерной графике и зрению. СПб: СПбГУ ИТМО, 2010. С. 315–319.
2. Лоханова А. И., Симончик К. К., Козлов А.В. Алгоритм детектирования музыкальных фрагментов в задачах речевой обработки // DSPA–2010. Тр. 12-й Межд. конф. «Цифровая обработка сигналов и ее применение». М., 2010. Т. 1. С. 210–213.
3. Идентификация дикторов на основе сравнения статистик основного тона голоса / С.Л. Коваль, П.В. Лабутин, Е.В. Малая и др. // Информатизация и информационная безопасность правоохранительных органов. Тр. XV Межд. науч. конф. М.: Академия управления МВД России, 2006. С. 324–327.
4. Comparison of Voice Activity Detection Algorithms for VoIP / R. Prasad et al. // ISCC'02. Proc. 7th IEEE Symposium on Computers and Communications. Washington: IEEE Computer Society, 2002. P. 530.
5. Симончик К. К., Галинина О. С., Капустин А. И. Алгоритм обнаружения речевой активности на основе статистик основного тона в задаче распознавания диктора // Научно-технические ведомости СПбГПУ. 2010. Т. 103. № 4. С. 18–23.
6. Пеховский Т. С., Шулипа А. К. Гибрид генеративных и дискриминативных моделей для задачи диаризации в коротком телефонном диалоге // SPECOM–2011. Proc. 14th Intern. Conf. «Speech and Computer». Kazan, 2011. P. 389–394.
7. Kenny P., Reynolds D., Castaldo F. Diarization of Telephone Conversations Using Factor Analysis // IEEE Journal of Selected Topics in Signal Processing. 2010. Vol. 4. No. 6. P. 1059–1070.

8. Koval S., Bekasova V., Khitrov M., Raev A. Pitch Detection Reliability Assessment for Responsible Applications // EUROSPEECH'97. Proc. 5th European Conf. on Speech Communication and Technology. Rhodes, 1997. P. 489–492.
9. Koval S. L. Formants Matching as a Robust Method for Forensic Speaker Identification // SPECOM'2006. Proc. XI Intern. Conf. «Speech and Computer». St. Petersburg, 2006. P. 125–128.
10. Капустин А. И., Симончик К. К. Система верификации дикторов по голосу на основе использования СГР–SVM подхода // DSPA–2010. Тр. 12-й Межд. конф. «Цифровая обработка сигналов и ее применение». М., 2010. Т. 1. С. 207–210.
11. Reynolds D. A., Quatieri T. F., Dunn R. B. Speaker Verification Using Adapted Gaussian Mixture Models // Digital Signal Processing. 2000. Vol. 10. No. 1–3. P. 19–41.
12. Kenny P., Boulianne G., Ouellet P., Dumouchel P. Joint Factor Analysis Versus Eigenchannels in Speaker Recognition // IEEE Transactions on Audio, Speech and Language Processing. 2007. Vol. 15. No. 4. P. 1435–1447.
13. Pekhovsky T., Oparin I. Maximum Likelihood Estimations for Session-Independent Speaker Modeling // SPECOM–2009. Proc. XIII Intern. Conf. «Speech and Computer». St.-Petersburg, 2009. P. 267–270.
14. Comparative Evaluation of Maximum a Posteriori Vector Quantization and Gaussian Mixture Models in Speaker Verification / T. Kinnunen, J. Saastamoinen, V. Hautamaki et al. // Pattern Recognition Letters. 2009. Vol. 30. P. 341–347.
15. Pekhovsky T., Lokhanova A. Variational Bayesian Model Selection for GMM–Speaker Verification Using Universal Background Model // INTERSPEECH–2011. Proc. 12th Annual Conf. Florence, 2011. P. 2705–2708.
16. Front-End Factor Analysis for Speaker Verification / N. Dehak, P. Kenny, R. Dehak et al. // IEEE Transactions on Audio, Speech and Language Processing. 2011. Vol. 19. No. 4. P. 788–798.

Статья поступила в редакцию 14.05.2012