УДК 004.89, 004.93

Т.В. Ермоленко, Н.С. Клименко

# Метод текстонезависимой идентификации диктора на базе данных моделей дикторов в виде древовидной структуры

Предложен метод идентификации диктора на основе гауссовых смесей и разбиения акустического пространства голоса диктора на множество классов, учитывающих фонетические особенности звуков речи. Использована база данных древовидной структуры для хранения моделей дикторов. Предложены процедуры структурирования базы данных, ускоряющих поиск.

In the paper a speaker identification method is proposed. It based on Gaussian mixture models and the partition of acoustic space of speaker's voice on the set of classes that take into account the phonetic characteristics of speech sounds. Such an approach has increased the efficiency of identification. The database tree structure is used for storage of speaker's models, proposed structuring procedures of database accelerate search operation.

Запропоновано метод ідентифікації диктора на основі гауссових сумішей і розбиття акустичного простору голосу на множину класів, що враховують фонетичні особливості звуків мовлення. Використано базу даних деревовидної структури для зберігання моделей дикторів. Запропоновано процедури структурування бази даних, що прискорюють пошук.

**Введение.** Сегодня активно ведется внедрение голосовой биометрии в многопользовательские автоматизированные системы различного спектра применения. Основное преимущество голосовой идентификации перед другими биометрическими системами заключается в возможности получения и передачи биометрических данных в центр управления доступом без применения специализированных и дорогих сканеров биометрической информации. Кроме того, процесс аутентификации не требует от пользователя непосредственного контакта с элементами пропускной системы, что открывает возможность проведения данной процедуры удаленно, например, через Интернет или сеть мобильной связи. Биометрическая голосовая аутентификация – эффективное средство, широко используемое для обеспечения безопасности данных.

Автоматические системы распознавания диктора делятся на системы идентификации и верификации. Идентификация представляет собой процесс сравнения речевого фрагмента с образами множества дикторов, зарегистрированных в системе. Пользователь, не зарегистрированный в системе, будет идентифицирован как диктор, модель которого является ближайшей к распознаваемой. В случае необходимости выявления незнакомого диктора в систему добавляется универсальная фоновая модель [1], созданная из совокупности характеристик всех зарегистрированных пользователей. Верификация диктора – это процесс принятия или отклонения факта принадлежности речевого фрагмента заявленному пользователю. В итоге будет принят только тот речевой сигнал, у которого значение вероятности соответствия с образом диктора будет не ниже порогового.

Таким образом, основное отличие между идентификацией и верификацией – количество альтернативных решений, которое при идентификации равно количеству множества образов дикторов, а при верификации – только принятию или отклонению, независимо от количества зарегистрированных пользователей. Исходя из этого следует отметить, что эффективность идентификации диктора часто уменьшается при увеличении множества моделей, тогда как эффективность верификации диктора близка к постоянной величине. Следовательно, идентификация по голосу предъявляет повышенные требования к разделимости моделей, в то время как для верификации необходима точность передачи индивидуальных характеристик каждого диктора в отдельности. Усложняет задачу ряд факторов: нестационарность произношения, эффект реверберации голоса, а также искажения и помехи в каналах связи.

Повышенный спрос на создание текстонезависимых систем вызван как простотой использования, так и необходимостью применения в правоохранительной сфере. Но эффективность и скорость распознавания таких систем на данный момент значительно уступает текстозависимым аналогам, не обеспечивая достаточно вы-

сокой надежности распознавания дикторов. Создание алгоритмов, обеспечивающих высокую точность текстонезависимой идентификации и сохраняющих при этом приемлемые показатели вычислительной трудоемкости, которые в современных условиях приближаются к возможности вычислений в реальном времени, – актуальная задача.

В статье приведено численное исследование модифицированного метода построения модели диктора, описанного в [2]. Модификация заключается в разделении акустического пространства голоса диктора на широкие фонетические классы (ШФК) путем предварительной текстонезависимой сегментации речевого сигнала с одновременной классификацией его сегментов.

**Методы построения признаковых описаний и принятия решения в задачах распознавания диктора**

Обе задачи распознавания по голосу опираются на признаковые описания – наборы структурированных акустических признаков, вычисляемых по речевому сигналу диктора. На основе признаковых описаний формируются модели диктора, структура таких моделей весьма разнообразна и напрямую зависит от используемых акустических характеристик и классификаторов.

Индивидуальность речи диктора формируется особенностями строения его речевого тракта и состоянием нервной системы, оказывающей непосредственное влияние на процесс артикуляционной деятельности. Акустические характеристики обязаны передавать данные особенности диктора, а также сочетать в себе устойчивость к искажениям разного рода и компактность представления для возможности быстрой обработки, хранения и сравнения эталонных значений.

Методы решения обеих задач используют спектральные акустические признаки речевого сигнала на основе преобразований Фурье и вейвлет-спектра, кепстральных коэффициентов, а также их производных по времени в виде векторов действительных чисел. К наиболее часто используемым акустическим признакам можно отнести:

- мел-частотные кепстральные коэффициенты (*Mel Frequency Cepstral Coefficient – MFCC*);
- перцептуальные и кепстральные коэффициенты линейного предсказания.

Для учета динамической составляющей векторы моментальных характеристик, вычисленные на наборе последовательных окон, могут быть представлены в виде матрицы [3].

Перед формированием набора векторов, характеризующих особенности речи говорящего, для обеспечения повышения робастности систем идентификации диктора речевой сигнал подвергается предварительной обработке (подавление шума, нормализация).

Из методов классификации моделей наиболее распространены: векторное квантование, гауссовские смеси и метод опорных векторов.

Метод *векторного квантования* решает задачу разделением всего пространства признаков на области, в которых сконцентрированы акустические признаки диктора. Данный метод строит модель в виде набора векторов признаков, являющихся центроидами кластеров, не пересекающихся между собой. Структура модели может быть дополнена весовыми коэффициентами для усиления важности отдельных кластеров. Также возможна структура с ведением единой общей карты кластеров, по которой модели дикторов описываются кодовыми книгами – статистическими данными о вхождении векторов-признаков в кластеры фоновой модели.

Модели, создаваемые на основе *гауссовых смесей*, продолжают идею векторного квантования, но с той разницей, что классы в пространстве признаков описываются в виде многомерного вероятностного распределения. Основная идея – представить его в виде взвешенной суммы $M$ нормальных распределений:

$$p\left(\overline{x}\middle|\lambda\right)=\sum_{i=1}^{M} w_i p_i\left(\overline{x}\right),$$

где $\overline{x}$ – $N$-мерный вектор признаков; $w_i$ – веса компонентов модели; $p_i$ – многомерные функции плотности распределения составляющих модели.

Таким образом, полностью модель описывается векторами математического ожидания, ко-

вариационными матрицами и весами смесей для каждого компонента модели.

Широко используемым методом оценки параметров модели есть метод максимизации правдоподобия. Функция максимального правдоподобия имеет вид:

$$p\left(X\left|\lambda\right.\right)=\prod_{t=1}^{T}p\left(\overline{x}_t\left|\lambda\right.\right), \qquad (1)$$

где $X=\left\{\overline{x}_1,...,\overline{x}_T\right\}$ – последовательность векторов признаков.

Поскольку (1) – нелинейная функция от параметров модели, то ее непосредственное вычисление невозможно, поэтому оценки параметров могут быть получены итерационно. Исходя из предположения, что все дикторы одинаково вероятны, упрощенное правило классификации имеет вид:

$$p\left(X\left|\lambda_k\right.\right), \qquad (2)$$

где $S$ – количество дикторов.

При помощи метода *опорных векторов* в многомерном пространстве признаков определяется расположение гиперплоскости, являющейся равноудаленной от крайних (опорных) векторов противоположных классов. Таким образом можно выполнить разделение только двух акустических классов, а для большего множества дикторов может быть использована схема «один против каждого». В этом случае модель диктора состоит из множества гиперплоскостей, каждая из которых отделяет признаки данного диктора от остальных. Это означает, что для системы, состоящей из $N$ моделей дикторов, необходимо построение матрицы попарно разделяющих гиперплоскостей размерностью $N$x$N$. Альтернатива – обратное решение данной задачи, когда близкие модели дикторов объединяются в группы. Процедура группировки повторяется итерационно, а результат – древовидная структура, в узлах которой находятся уравнения, разделяющие ближайшие классы групповых признаков, а листья представляют собой непосредственно акустические признаки дикторов. При больших количествах дикторов данное представление значительно уменьшает объем хранимых данных и вычислительную сложность при распознавании. Эффективность будет зависеть от сбалансированности бинарного дерева. Чаще всего применяется схема «один против всех», цель которой – отделение признаков конкретного диктора от всех остальных. Решение в таком виде идеально приспособлено для задачи верификации, но также широко применяется и при идентификации диктора [4]. В случае линейной неразделимости для построения гиперплоскости между частично пересекающимися классами, ограничения дополняются скалярным параметром допуска. Другой способ, позволяющий распознавать линейно-неразделимые классы, – отображение исходного пространства признаков в пространство большей размерности, в котором классы могут быть разделены линейно. Данное преобразование выполняется с помощью функции ядра. Параметры метода (скалярный параметр допуска и параметры ядра), как правило, определяют с помощью перебора некоторого множества значений.

В течение последних нескольких лет смеси гауссовых моделей стали доминирующим подходом для моделирования в текстонезависимых приложениях распознавания диктора. Это доказано многочисленными исследованиями и описано в статьях, изданных на международных конференциях, таких как международная конференция по акустике речи и обработке сигналов *ICASSP*, *EuroSpeech*, *ICSLP* и т.д., а также статьями в трудах *ISCA* и *IEEE*.

В статье предлагается модификация метода идентификации на основе гауссовых смесей, использующего базу данных в виде древовидной (иерархической) структуры.

**Организация иерархической базы данных моделей дикторов для работы в режиме реального времени**

Система идентификации диктора как в режиме обучения, так и в режиме распознавания обращается к базе данных моделей речи дикторов. Идентификация диктора сводится к поиску в базе данных наиболее близкой модели, обучение – к добавлению новых записей. Следовательно, возникает необходимость в разработке структуры базы данных, обеспечивающей быстрый поиск и обучение системы.

Быстрый поиск обеспечивает процедура индексирования данных, наиболее популярные подходы построения индекса – хеш-таблицы и деревья. Преимущество хеш-таблиц – отсутствие необходимости в логической упорядоченности значений ключей физических записей. Эффективность доступа зависит от распределения ключей, алгоритма их преобразования (хеш-функции) и распределения памяти. Они позволяют находить необходимую запись по ключам ассоциативного массива, получаемых сверткой исходного признака к битовой последовательности фиксированного размера. Но в задачах идентификации диктора большая размерность вектора признаков сильно усложняет построение хеш-функции с пригодным для применения уровнем коллизий и размером хеш-суммы.

В связи с этим для организации поиска по базе данных моделей дикторов в режиме реального времени выбрана структура хранения моделей в виде дерева с итеративным обобщением моделей. Листом дерева есть модель одного диктора, узлом – фоновая модель речи некоторой группы дикторов, а корнем – обобщенная или универсальная фоновая модель. Получаемое в процессе добавления моделей дерево не является бинарным и сбалансированным, поэтому его структура и скорость поиска напрямую зависят от взаимного расположения моделей в акустическом пространстве. Для балансировки проводится процедура оптимизации.

*Процедура поиска* выполняется от корня дерева и продвигается в глубину. На каждом этапе вычисляется вероятность принадлежности набора векторов признаков (ВП) к выбранной модели. Если значение выше порогового и текущая позиция – лист, то поиск считается успешно завершенным. Если среди соседей не найдена модель, соответствующая набору ВП, то поиск завершается, и делается вывод об отсутствии модели в дереве или о неполноте идентификационных данных (если процедура завершена на листьях). Таким образом, одной процедурой решаются задачи идентификации и верификации диктора.

*Процедура добавления* новых записей, необходимая при обучении системы, выполняется аналогично поиску в глубину наиболее подходящей модели: на каждом шаге определяется узел, потомком которого есть добавляемая модель. Поиск прекращается при уменьшении значения апостериорной вероятности принадлежности добавляемой модели текущему узлу. Если поиск достиг листа, то на его месте строится узел, иначе добавление выполняется к текущему узлу. На завершающем этапе пересчитываются все фоновые модели – предки добавленного листа.

Для предотвращения неверного структурирования, возможного после добавления нескольких записей в базу данных, и с целью повышения эффективности идентификации реализована *процедура оптимизации* (*корректировки*) структуры базы данных. Она заключается в обработке всего множества моделей и требует значительных вычислительных ресурсов. Но необходимость выполнения данной процедуры уменьшается с увеличением количества моделей, поскольку вероятность изменения параметров фоновых моделей убывает.

Начало оптимизации – рекурсивная процедура кластеризации всего признакового пространства, т.е. наиболее ресурсоемкий этап задачи. Результат – набор из $N$ кластеров, удовлетворяющих условию максимального межкластерного разброса. $N$ находится в пределах от двух до пяти, что необходимо для предотвращения чрезмерного ветвления дерева и затруднения принятия решения на каждом узле сравнения. На каждом уровне иерархии проверяется, есть ли смещение центроидов кластеров до и после корректировки. Если смещение произошло, то фоновая модель соответствующего узла и множество ее потомков пересчитываются.

Процедуру оптимизации имеет смысл выполнять после добавления большого количества новых записей (более 20 процентов от количества занесенных в базу данных моделей речи дикторов). Как показали численные исследования, процедура оптимизации может не только увеличить эффективность идентификации за счет корректировки состава кластеров призна-

кового пространства, но уменьшить глубину дерева, а следовательно, увеличить скорость поиска по нему. Так, при добавлении до 100 моделей к существующей базе данных из 75 моделей речи дикторов глубина дерева уменьшилась с 14 до 11 уровней, увеличив тем самым среднюю скорость поиска на восемь процентов.

**Исследование эффективности модифицированного метода идентификации по голосу, учитывающего широкие фонетические классы звуков**

При разделении признакового пространства голоса диктора на ШФК становится возможным формирование модели диктора по акустическим признакам звуков, близких по способу формирования, что позволит создать более точную модель речи диктора. В статье использовались четыре класса звуков русского языка:

- *Voc* – гласные {[і], [е], [о], [у], [а], [и]};
- *Sh* – глухие согласные {[ф], [с], [х], [ш], [ф’], [с’], [х’], [щ], [ц], [ч]};
- *Cons* – звонкие согласные {[в], [з], [ж], [в'], [з'], [ж'], [б], [д], [г], [б'], [д '], [г']};
- *Son* – сонорные {[й], [л], [л '], [м], [н], [м'], [н '], [р], [р']}.

Кроме звуков речи, в качестве пятого класса был использован шум – фрагмент сигнала, не содержащий речь.

Для идентификации применялся метод гауссовых смесей, в качестве акустических характеристик использованы *MFCC*, формирующие 13-мерный вектор признаков, этого количества коэффициентов достаточно для обработки речевых сигналов.

Речевой сигнал разбивался на фреймы длиной около 20 мс с половинным перекрытием, далее проводилась процедура классификации фреймов по ШФК. Модели ШФК, по которым осуществлялась классификация, строились на основе гауссовых смесей размерностью 10, для обучения были использованы записи дикторов (мужчин и женщин с различными голосовыми данными) общей продолжительностью около 20 мин.

По набору ВП, полученных на множестве фреймов, принадлежащих одному ШФК, выполнялась кластеризация методом *K*-средних с итеративным добавлением центроидов (разделением кластера с максимальным радиусом на два). Количество центроидов, а следовательно, и размерность гауссовой смеси определялось согласно критерию эффективности описания выборки смесью из $k$ компонент, включающему в себя штраф на количество компонент (критерий *ICL–BIC*), описанному в [5]. Для окончательного позиционирования центроидов применялся метод максимизации правдоподобия. Создание модели диктора, соответствующей определенному ШФК, завершалось построением гауссовой смеси с использованием полученных центроидов. Результирующая модель представляет собой набор из четырех моделей диктора, сформированных для различных ШФК:

$$\lambda_k = \left(\lambda_k^{Voc}, \lambda_k^{Sh}, \lambda_k^{Cons}, \lambda_k^{Son}\right).$$

В численном исследовании эффективности использования комплексной модели диктора принимали участие 100 дикторов с различными голосовыми данными. Для построения моделей записаны фрагменты речи дикторов средней продолжительностью одна минута. Запись осуществлялась динамическим микрофоном в помещении без посторонних шумов (уровень шума –45 *dB*) с частотой дискретизации 44,1 кГц и глубиной квантования 16 бит.

Для проведения сравнительного анализа эффективности идентификации согласно методу, изложенному в [2], были получены модели дикторов, не учитывающие разделение по ШФК, а также модели, обученные только на фреймах, принадлежащих одному ШФК.

Одна из проблем при обучении смесей гауссовых моделей – выбор числа компонентов модели. Авторами статьи использовался критерий эффективности *ICL–BIC*. Максимальные и минимальные размерности сформированных моделей дикторов, полученных в результате численных исследований с помощью данного критерия, представлены в табл. 1.

При создании моделей предельной выбрана размерность 20. Как видно из результатов моделирования без учета ШФК, данное значение достигнуто при обработке образцов голоса каждого диктора. Этот факт свидетельствует о большой энтропии кластеризируемых данных.

При построении моделей для каждого ШФК ситуация чрезмерной кластеризации наблюдалась только для класса *Voc*.

**Таблица 1**. Максимальные и минимальные количества гауссовых компонент в смеси, полученных при построении моделей дикторов по критерию эффективности *ICL–BIC*

| Тип модели | *MIN* | *MAX* |
|---|---|---|
| Без учета ШФК | 20 | 20 |
| *Voc* | 11 | 20 |
| *Sh* | 8 | 19 |
| *Cons* | 6 | 11 |
| *Son* | 16 | 20 |

Отметим, что выполнено построение моделей с предельной размерностью большего порядка. Результатом стало как увеличение средней размерности до 27, так и проявление свойств чрезмерной кластеризации вследствие избыточного разделения областей с большим внутрикластерным разбросом. Поэтому было принято решение не повышать размерность.

В статье принадлежность последовательности ВП модели диктора определялась с помощью функции (3), являющейся аналогом функции правдоподобия (1), в которой с целью сглаживания эффекта, производимого близкими к нулю оценками вероятности, вместо произведения использовалась усредненная сумма:

$$F(k) = \frac{\sum_{t=1}^{T} p\left(\overline{x}_t \middle| \lambda_k\right)}{T}, \qquad (3)$$

где $T$ – количество фреймов речевого сигнала, по которому проводится идентификация.

В этом случае правило классификации принимает вид: $S = \arg\max_{1 \le k \le 100} F(k)$.

Значение функции (3) должно быть максимальным, если модель и речевой сигнал, по которому получена последовательность ВП, принадлежат одному диктору. Данная зависимость демонстрирует точность передачи характеристик диктора, а низкий разброс получаемых значений может свидетельствовать о стабильности результатов. Проведено сравнение последовательности векторов признаков речевых фрагментов каждого диктора с соответствующей ему моделью. Акустические признаки получены из сигналов продолжительностью не менее пяти с, а по значениям функции (3) вычислены математическое ожидание (МО) и среднеквадратичное отклонение (СКО).

При рассмотрении полученных значений функции (3), соответствующих моделям без учета ШФК, наблюдался самый высокий разброс результатов. По сравнению с этими данными МО комплексных моделей выросло в среднем по всем дикторам в пять раз, в то время как СКО снизилось на 15 процентов.

При проведении сравнительного анализа моделей, обученных на фреймах, принадлежащих одному ШФК, и моделей без учета ШФК можно сказать следующее:

- значения МО и СКО функции (3) для моделей, обученных на фреймах класса *Voc*, сравнимы с показателями моделей без учета ШФК;
- для моделей, обученных на фреймах классов *Sh* и *Cons*, МО значений функции (3) возросло в среднем по всем дикторам в два раза, однако их СКО сравнимо с СКО для моделей без учета ШФК;
- исследуемые показатели для моделей, обученных на фреймах класса *Son*, – наилучшие.

Проведем аналогичное исследование поведения функции (3) для случая, когда модель и сигнал, подлежащий идентификации, принадлежат разным дикторам. В данном случае значение функции (3) должно стремиться к нулю.

Полученные значения в 10–50 раз меньше, чем значения, вычисленные для случая, когда модель и сигнал, по которому проводится идентификация, принадлежат одному диктору. Соответствующие данные приведены в табл. 2.

**Таблица 2**. Отношение статистических параметров значений функции (3) различных типов моделей к значениям, соответствующим параметрам моделей без разделения на ШФК

| Тип модели | МО, % | СКО, % | МО, % | СКО, % |
|---|---|---|---|---|
| | модель и сигнал, подлежащий идентификации, принадлежат одному диктору | | модель и сигнал, подлежащий идентификации, принадлежат разным дикторам | |
| Комплексная модель | 508,6 | 74,1 | 97,7 | 144,3 |
| *Voc* | 87,3 | 93,7 | 99,2 | 118,3 |
| *Sh* | 192,4 | 104 | 101 | 214,2 |
| *Cons* | 201,6 | 86,3 | 95,1 | 396,7 |
| *Son* | 1088 | 107,1 | 98,4 | 483,2 |

При проведении идентификации с помощью различных моделей дикторов результаты показали перспективность применения комплексной модели (табл. 3).

**Таблица 3**. Показатели эффективности идентификации при использовании различных типов моделей диктора

| Тип модели | Вероятность идентификации, % |
|---|---|
| Без учета ШФК | 94,8 |
| Комплексная модель | 98,7 |
| *Voc* | 94,3 |
| *Sh* | 96,7 |
| *Cons* | 95,1 |
| *Son* | 97,2 |

**Заключение**. Проведено исследование эффективности метода идентификации, использующего комплексную модель диктора, учитывающую ШФК. Анализируя полученные результаты, можно сделать следующие выводы.

Наиболее подходящей структурой для организации поиска по базе данных моделей дикторов в режиме реального времени есть древовидная модель с итеративным обобщением моделей. Листом дерева является модель речи диктора, узлом – фоновая модель речи некоторой группы дикторов, а корнем – обобщенная или универсальная фоновая модель.

Использование комплексной модели диктора, элементы которой получены в результате обработки участков сигнала, принадлежащих различным ШФК, позволило повысить вероятность идентификации более чем на 3 процента.

Установлено, что элементы комплексной модели диктора, обученные на фреймах, принадлежащих только одному ШФК, обладают различными разделительными способностями в зависимости от состава фонетического класса, а значит, вносят разный вклад в идентификационные свойства результирующей модели.

Эффективность идентификации диктора можно повысить за счет:

- улучшения разделяющих свойств ВП путем добавления к нему робастных акустических характеристик, обладающих идентификационными свойствами;
- добавления в целевую функцию решающего правила весовых коэффициентов, отражающих разделительные способности моделей, обученных на одном ШФК.

Эффективность верификации диктора повышается путем:

- точности промежуточных и общей фоновых моделей при помощи метода опорных векторов;
- использования отдельного множества решающих правил по схеме «один против всех».

Сравнительное исследование различных моделей формирования признакового описания речевых сигналов и систем распознавания дикторов – цель определения перспективных направлений их создания.


1. *Wei-Qiang Zhang, Jia Liu.* Discriminative Universal Background Model Training for Speaker Recognition // Speech and Language Technologies. – 2011. – N 6. – P. 241–256.
2. *Садыхов Р.Х., Ракуш В.В.* Модели гауссовых смесей для верификации диктора по произвольной речи // Докл. Белорусского гос. ун-та информ. и радиоэл. – Минск, 2003. – № 4. – С. 95–103.
3. *Wu Q., Zhang L.Q., Shi G.C.* Robust feature extraction for speaker recognition based on constrained nonnegative tensor factorization // J. of comp. science and technology. – 2010. – N 25(**6**). – P. 745–754.
4. *Bartlett P., Shawe-Taylor J.* Generalization performance of support vector machines and other pattern classifiers // Advances in Kernel Methods. – Cambridge: MIT Press, 1998. – P. 43–54.
5. *Сорокин В.Н., Цыплихин А.И.* Верификация диктора по спектрально-временным параметрам речевого сигнала // Информационные процессы. – 2010. – Т. 10, № 2. – С. 87–104.



*E-mail*: *naturewild*71*@gmail.com*, *nk@xaker.ru*