УДК 004.89, 004.93

Т.В. Ермоленко, Н.С. Клименко

ИСПОЛЬЗОВАНИЕ СЕГМЕНТАЦИИ РЕЧЕВОГО СИГНАЛА ДЛЯ ПОСТРОЕНИЯ КОМПЛЕКСНОЙ МОДЕЛИ ДИКТОРА В СИСТЕМЕ ИДЕНТИФИКАЦИИ ГОВОРЯЩЕГО

Ермоленко Т.В., Клименко Н.С. **Использование сегментации речевого сигнала для построения комплексной модели диктора в системе идентификации говорящего.**
Аннотация. Статья посвящена разработке комплексной модели диктора в задаче текстонезависимой идентификации по голосу. Комплексная модель базируется на методе гауссовых смесей. Ее формируют по речевому сигналу, который предварительно сегментируется на фрагменты, соответствующие различным фонетическим классам звуков. Предложен способ структурирования моделей дикторов. Модели дикторов структурированы в виде дерева, что позволило проводить идентификацию диктора без выполнения полного перебора всего множества моделей. Проведенные исследования показали, что деление акустического пространства голоса диктора на множество классов, представляющих некоторые фонетические события, приводит к увеличению эффективности идентификации по голосу, а предложенное структурирование множества моделей дикторов ускоряет операцию поиска.
Ключевые слова: кластеризация, гауссовы смеси, модели дикторов, широкие фонетические классы, мел-частотные кепстральные коэффициенты.

Yermolenko T.V., Klymenko M.S. **Usage of Speech Signal Segmentation for the Construction of Complex Model in the Speaker Identification System.**
Abstract. The article is devoted to development of a complex speaker model for using at the text-independent speaker identification. The complex speaker model is based on gaussian mixture method. The model is formed by preliminary segmented speech signal, where each segment matches to certain broad phonetic class. Method of speaker models structuring is proposed. Speaker models are structured as a tree, which allows to identify speaker without running a full search on the set of models. Researches have shown the division of the acoustic space of speaker's voice on the set of classes that represent some phonetic events, increases the efficiency of voice identification and the proposed structuring method of models accelerates the search operation.
Keywords: clustering, gaussian mixture, speaker models, broad phonetic classes, mel-frequency cepstral coefficients.

1. Введение. Биометрические данные о человеке в настоящее время находят свое применение в широком спектре задач, становясь частью не только профессиональных автоматизированных систем, но и бытовых устройств, ориентированных на многопользовательский режим работы. Основное преимущество голосовой идентификации перед рядом других биометрических параметров заключается в возможности получения и передачи биометрических данных в блок управления доступом без применения специализированных и дорогостоящих 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.

Подавляющее большинство систем идентификации по голосу являются текстозависимыми. Они демонстрируют высокую эффективность (точность и скорость) идентификации по произнесениям слов из заданного набора. Текстонезависимые же системы на данный момент значительно уступают им по этим показателям. При этом текстонезависимые системы идентификации являются более перспективными в связи с простотой использования с точки зрения пользователя в комплексах интеллектуального управления или анализа ситуации [1]. Особенно актуально применение текстонезависимых систем идентификации в правоохранительной сфере для автоматизации проведения фоноскопических экспертиз. Поэтому создание алгоритмов текстонезависимой идентификации, обеспечивающих высокую точность и приемлемую вычислительную сложность, является актуальной задачей.

Идентификация говорящего является одной из задач распознавания диктора. К этому классу задач также относится верификация диктора. Тогда как при верификации произнесение сопоставляется с образом одного диктора, идентификация предполагает сопоставление произнесения со всем множеством имеющихся в системе образов дикторов. То есть эффективность идентификации диктора значительно падает с ростом множества дикторов.

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. От выбора структуры модели, акустических характеристик и классификаторов зависит эффективность системы. При этом задача идентификации предъявляет повышенные требования к разделимости моделей.

Целью данной работы является построение комплексной модели диктора, применимой в системе текстонезависимой идентификации диктора. Для достижения цели в работе решены следующие задачи: выполнен анализ методов построения признаковых описаний и принятия решения в задачах распознавания диктора; предложена комплексная модель диктора и исследована её эффективность; разработан способ организации базы данных моделей дикторов,

позволяющий идентифицировать диктора в режиме реального времени.

2. Выбор методов построения признаковых описаний и принятия решения в задачах распознавания диктора. Индивидуальность речи диктора формируется особенностями строения его речевого тракта и состоянием нервной системы, которая оказывает непосредственное влияние на процесс артикуляционной деятельности. Выбранные акустические характеристики должны передавать данные особенности диктора, а также сочетать в себе устойчивость к искажениям разного рода (нестационарность произношения, эффект реверберации голоса, искажения и помехи в каналах связи) и компактность представления для возможности быстрой обработки, хранения и сравнения эталонных значений.

В настоящее время 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);

– линейно-частотные кепстральные коэффициенты;

– перцептуальные коэффициенты линейного предсказания.

Для учета динамической составляющей векторы моментальных характеристик, вычисленные на наборе последовательных окон, могут быть представлены в виде матрицы [4]. Более распространенным способом является добавление к вектору признаков (ВП) его первой и второй производной по времени, расширяя тем самым область анализа в 3 раза. Это может быть полезно при поиске в сигнале коротких смычных или взрывных звуков речи, которые могут полностью уместиться в данный временной интервал.

Из методов классификации моделей наиболее распространенными на данный момент являются: векторное квантование, гауссовы смеси, скрытые марковские модели 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. Объем информации, необходимой для описания модели таким методом, компактен, а на качество идентификации сильно влияет степень детализации. По сравнению с другими методами, высока вероятность ошибки 1-го рода, поэтому векторное квантование целесообразно использовать в задаче идентификации, но не верификации.

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, избавляясь тем самым от недостатка детализации квантования. Основная идея — представить его в виде взвешенной суммы M нормальных распределений:

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

где $\overline{x}$ – N-мерный вектор признаков;

w_i – веса компонентов модели;

p_i – многомерные функции плотности распределения составляющих модели.

Таким образом, модель описывается векторами математического ожидания, ковариационными матрицами и весами смесей для каждого компонента модели.

Широко используемым способом оценки параметров модели является метод максимизации правдоподобия, функция которого имеет вид:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\overline{x}_t \mid \lambda), \tag{1}$$

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

Исходя из предположения, что все дикторы одинаково вероятны, упрощенное правило классификации имеет вид:

$$res = \arg\max_{1 \le k \le S} p\left(X \mid \lambda_k\right), \tag{2}$$

где S – количество дикторов.

Метод скрытых марковских моделей (СММ) определяет модель последовательностью состояний. Каждому состоянию соответствует

распределение вероятностей появления в данном состоянии, а также матрица вероятностей переходов [5]. Поскольку идентификационные характеристики диктора чаще всего скрыты в кратковременных участках фонем и межфонемных переходов, то необходимость использования цепочек состояний отсутствует. Таким образом, данный метод редко используется в задачах идентификации, но находит широкое применение в задачах распознавания речи [6], т. к. цепочками состояний возможно описать как модели фонем, так и их последовательности (слова) и целые предложения, выходя на синтаксический уровень. Гауссовы смеси представляются упрощенной реализацией СММ, в которой учитывается единственное состояние. Однако данный недостаток гауссовых смесей может быть компенсирован добавлением к вектору признаков его производных по времени.

Метод опорных векторов позволяет определить в многомерном пространстве признаков расположение гиперплоскости, являющейся равноудаленной от крайних (опорных) векторов двух противоположных классов. 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 одного из остальных. Это означает, что для системы, состоящей из N моделей дикторов, необходимо построение матрицы попарно разделяющих гиперплоскостей размерностью NxN. Широко распространена альтернативная схема «один против всех», целью которой является отделение признаков конкретного диктора от всех остальных. Решение в таком виде идеально приспособлено для задачи верификации, но также широко применяется и при идентификации диктора [7].

В случае линейной неразделимости для построения гиперплоскости между частично пересекающимися классами ограничения дополняются скалярным параметром допуска. Другим способом, позволяющим распознавать линейно-неразделимые классы, является 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 модели гауссовых смесей стали доминирующим подходом в текстонезависимых приложениях

распознавания диктора. Это доказано многочисленными исследованиями и описано в статьях, изданных в сборниках трудов международных конференций, таких как международная конференция по акустике речи и обработке сигналов ICASSP, EuroSpeech, ICSLP и т.п., а также статьями в Трудах ESCA и Трудах IEEE.

В данной работе в качестве методов построения признаковых описаний выбрано MFCC, а в качестве метода принятия решения — гауссовы смеси.

Поскольку акустические признаки в различной степени характеризуют индивидуальность речи диктора, то для повышения качества идентификации часто используют несколько признаков, получаемых на основе различных представлений сигнала. Например, существуют системы идентификации, использующие одновременно мел-частотные кепстральные коэффициенты и линейно-частотные кепстральные коэффициенты.

Совместное использование признаков возможно несколькими способами. Объединение всех признаков в вектор приводит к заметному увеличению объема вычислений, но после проведения правильного взвешивания элементов вектора можно достичь максимального результата. Вторым подходом является создание набора независимых систем, каждая из которых работает с использованием своих акустических признаков, а окончательное решение принимается методом взвешенного голосования. Наиболее перспективным является метод бустинга, при помощи которого строится сильный классификатор на основании множества более слабых, различных как по акустическим признакам, так и по методу классификации.

В работе предложено дополнительно к данному решению использовать отдельные смеси для групп звуков речи, отстоящих друг от друга в пространстве акустических признаков. Эти группы звуков будем называть широкими фонетическими классами (ШФК). В данной работе создана комплексная модель диктора и исследована эффективность её использования в задаче идентификации. Точность идентификации диктора на основе предложенной модели будем сравнивать с точностью идентификации на основе метода построения модели диктора [2].

Идентификация диктора в рамках данного подхода предполагает выполнение предварительной текстонезависимой сегментации речевого сигнала с одновременной классификацией его сегментов. Модель диктора описывается набором моделей, полученных в результате обработки акустических характеристик на участках сигнала,

соответствующих разным ШФК.

3. Создание модели диктора на сегментированном речевом сигнале. Исследование ее эффективности. Комплексная модель диктора представляет собой набор моделей диктора, сформированных для различных ШФК.

Разделение признакового пространства голоса диктора на ШФК делает возможным вычисление модели по акустическим признакам звуков, близких по способу формирования. Созданные таким образом модели в совокупности позволяют создать более точную модель речи диктора.

Для создания комплексной модели речевой сигнал был предварительно сегментирован на участки, принадлежащие различным ШФК. Для этого он 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, в качестве ВП использовались MFCC, обучение проводилось на записях дикторов (мужчин и женщин с различными голосовыми данными) общей продолжительностью около 20 минут.

По набору ВП, полученных на множестве фреймов, принадлежащих одному ШФК, выполнялась кластеризация методом К-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 [8]). Для окончательного позиционирования центроидов применялся метод максимизации правдоподобия. Создание модели диктора, соответствующей определенному ШФК, завершалось построением гауссовой смеси с использованием полученных центроидов.

При проведении численных исследований использовались 4 класса звуков русской речи:

– *Voc* – гласные {[і], [е], [о], [у], [а], [и]};

– *Sh* – глухие согласные {[ф], [с], [х], [ш], [ф’], [с’], [х’], [щ], [ц], [ч]};

– *Con* – звонкие согласные {[в], [з], [ж], [в'], [з'], [ж'], [б], [д], [г], [б'], [д '], [г']};

– *Son* – сонорные {[й], [л], [л '], [м], [м'], [н], [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.

Таким образом, результирующая модель представляет собой набор из 4 моделей диктора, сформированных для различных ШФК:

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

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

Для проведения сравнительного анализа был реализован метод, изложенный в [2], и получены модели дикторов, которые не учитывают разделение по ШФК. Кроме того, исследовались модели, обученные только на фреймах, принадлежащих одному ШФК.

Одной из проблем при обучении смесей гауссовых моделей является выбор числа компонентов модели. В данной работе авторами использовался критерий эффективности ICL-BIC. При создании моделей предельной была выбрана размерность 20. Также было выполнено построение моделей с предельной размерностью большего порядка. Результатом стало как незначительное увеличение средней размерности, так и проявление свойств чрезмерной кластеризации вследствие избыточного разделения областей с большим внутрикластерным разбросом. Поэтому было принято решение не повышать размерность.

В данной работе 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:

$$F(k) = \frac{\sum_{t=1}^{T} p(\overline{x}_t \mid \lambda_k)}{T}, \tag{3}$$

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. Было проведено сравнение последовательности ВП, вычисленных по речевым фрагментам каждого диктора, с соответствующей ему моделью. Акустические признаки получены из сигналов продолжительностью не менее 5 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 5 раз, в то время как СКО снизилось на 15%.

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 2 раза, однако их СКО сравнимо с СКО для моделей без учета ШФК;

– исследуемые показатели для моделей, обученных на фреймах класса *Son*, являются наилучшими.

В случае, когда модель и сигнал, подлежащий идентификации, принадлежат разным дикторам, значение функции (3) должно стремиться к 0. Соответствующие данные приведены в таблице 1. Как видно из таблицы 1, полученные значения (3) в 10 – 50 раз меньше, чем значения, вычисленные для случая, когда модель и сигнал, по которому проводится идентификация, принадлежат одному диктору.

Таблица 1. **Отношение статистических параметров значений функции (3) различных типов моделей к значениям, соответствующим параметрам моделей без разделения на ШФК**

Тип модели	МО, %	СКО, %	МО, %	СКО, %
	идентифицируемый сигнал и модель принадлежат одному диктору		идентифицируемый сигнал и модель 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. 2).

Таблица 2. **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

4. Организация иерархической базы данных моделей дикторов для работы в режиме реального времени. Работа системы идентификации диктора может происходить в режимах обучения и распознавания: при обучении происходит добавление или модификация записей в базе данных моделей речи дикторов, а идентификация диктора сводится к поиску в базе данных наиболее близкой модели. Следовательно, возникает необходимость в разработке структуры базы данных, обеспечивающей быстрый поиск и обучение системы.

Скорость поиска обеспечивается за счет проведения

индексирования данных, наиболее популярными подходами построения индекса являются 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 задачу построения 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 была выбрана структура хранения моделей в виде дерева с итеративным обобщением моделей, пример схемы которой изображен на рисунке 1.

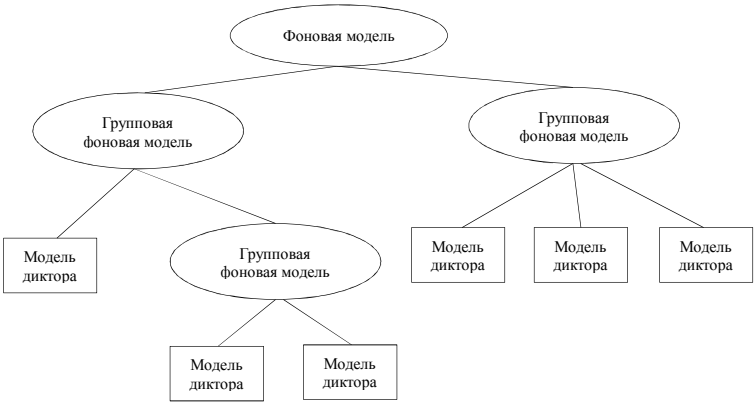

Рис. 1. Пример древовидного структурирования базы данных моделей.

Древовидная структура может быть сохранена в реляционной базе данных без значительных накладных затрат: индексируется лишь поле уникального идентификатора моделей, а для ведения иерархии с каждой моделью хранится номер ее родительского элемента. Листом дерева является 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 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 является 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 может быть выполнен пересчет всех фоновых моделей — предков добавленного листа. Данный шаг может быть пропущен в силу того, что с учетом имеющихся вычислительных мощностей и полученной глубины дерева невозможно успеть провести вычисление в заданный лимит времени.

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.

Началом оптимизации является рекурсивная процедура кластеризации всего признакового пространства — наиболее ресурсоемкий этап задачи. Результатом является набор из N кластеров, удовлетворяющих условию максимального межкластерного разброса. N находится в пределах от 2 до 5, 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 целесообразно выполнять после добавления большого количества новых записей (более 20% 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 признакового пространства, но и 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 8%.

5. Выводы. В данной работе было проведено исследование эффективности метода идентификации, использующего комплексную модель диктора, которая учитывает ШФК. Анализируя полученные результаты, можно сделать следующие выводы.

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

2. В результате проведения численных исследований 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.

3. Наиболее подходящей структурой для организации поиска по базе данных моделей дикторов в режиме реального времени является 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. Для решения задач идентификации и верификации диктора была разработана процедура поиска в глубину, с целью повышения эффективности идентификации после добавления значительного количества новых записей в базу данных — процедура оптимизации (корректировки) структуры базы данных. Как показали численные исследования, процедура оптимизации может не только увеличить эффективность идентификации за счет корректировки

состава кластеров признакового пространства, но и уменьшить глубину дерева, следовательно, увеличить скорость поиска по нему.

4. 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 можно повысить за счет:

– повышения точности промежуточных и общей фоновых моделей при помощи метода опорных векторов;

– использования отдельного множества решающих правил по схеме «один против всех».

Литература

1. *Ронжин Ал.Л., Будков В.Ю., Ронжин Ан.Л.* Формирование профиля пользователя на основе аудиовизуального анализа ситуации в интеллектуальном зале совещаний // Труды СПИИРАН. 2012. Вып. 23. С. 482–494.
2. *Садыхов Р.Х., Ракуш В.В.* Модели гауссовых смесей для верификации диктора по произвольной речи // Доклады БГУИР. 2003. №4. С. 95–103.
3. *Wei-Qiang Zhang, Jia Liu* Discriminative universal background model training for speaker recognition // Speech and Language Technologies, 2011. P. 241–256.
4. *Wu Q, Zhang L.Q., Shi G.C.* Robust feature extraction for speaker recognition based on constrained nonnegative tensor factorization // Journal of computer science and technology. №25(4). P. 745–754.
5. *Young S.G.* The HTK hidden markov model toolkit: design and philosophy // Cambridge university engineering dept. Technical report 1993.
6. *Кипяткова И.С., Карпов А.А.* Эксперименты по распознаванию слитной русской речи с использованием сверхбольшого словаря // Труды СПИИРАН. Вып. 12, СПб.: Наука, 2010. С. 63–74.
7. *Bartlett P., Shawe-Taylor J.* Generalization performance of support vector machines and other pattern classifiers // Advances in Kernel Methods. MIT Press, 1998. 13 p.
8. *Сорокин В.Н., Цыплихин А.И.* Верификация диктора по спектрально-временным параметрам речевого сигнала // Информационные процессы. Т. 10, № 2. С. 87–104.

Ермоленко Татьяна Владимировна — канд.техн.наук, начальник отдела распознавания речевых образов Института проблем искусственного интеллекта НАН Украины и МОН Украины, доцент кафедры Программного обеспечения интеллектуальных систем ДонНТУ. Область научных интересов: цифровая обработка сигналов, распознавание речи, идентификация диктора, автоматическая обработка естественно-языковых текстов. Число научных публикаций — 36. naturewid71@gmail.com; Украина, 83048, г. Донецк, ул. Артема, 118 б.

Yermolenko Tatyana Vladimirovna — PhD, senior researcher, head of Speech recognition department at the Institute of Artificial Intelligence, associate professor of Intelligent Systems

Software at DonNTU. Research interests: digital signal processing, speech recognition, speaker identification, natural language processing. The number of publications — 36. naturewid71@gmail.com; Ukraine, 83048, Donetsk, Artyoma str. 118 b.

Клименко Никита Сергеевич — младший научный сотрудник отдела распознавания речевых образов Института проблем искусственного интеллекта НАН Украины и МОН Украины. Область научных интересов: цифровая обработка сигналов, идентификация диктора, автоматическое распознавание речи. nk@xaker.ru; Украина, 83048, г. Донецк, ул. Артема, 118 б.

Клименко Никита Сергеевич — junior researcher at Speech recognition department of the Institute of Artificial Intelligence. Research interests: digital signal processing, speaker identification, speech recognition. nk@xaker.ru; Ukraine, 83048, Donetsk, Artyoma str. 118 b.

Рекомендовано лабораторией речевых и многомодальных интерфейсов, заведующий лабораторией Ронжин Ан.Л., д-р техн. наук, доцент.

Статья поступила в редакцию 01.03.2013.

РЕФЕРАТ

Ермоленко Т.В., Клименко Н.С. **Использование сегментации речевого сигнала для построения комплексной модели диктора в системе идентификации говорящего.**

В статье предложен подход к созданию комплексной модели диктора на основе гауссовых смесей. Элементы модели формируются по предварительно сегментированному речевому сигналу на участки, причем каждый сегмент соответствует определенному фонетическому классу.

Показано, что эффективность системы идентификации, выражающаяся в показателях точности и скорости, зависит от структуры модели диктора, выбранных акустических характеристик и классификаторов, способа структурирования множества моделей дикторов.

Анализ методов построения признаковых описаний и принятия решения в задачах распознавания диктора позволил выбрать в качестве таковых мел-кепстральные частотные коэффициенты и гауссовы смеси.

В качестве структуры модели диктора предложен набор гауссовых смесей, каждый из которых описывает определенный широкий фонетический класс. Построение моделей выполняется с помощью подсистемы предварительной дикторонезависимой сегментации речевого сигнала, выполняющей также классификацию сегментов.

Результаты численных исследований показали, что использование предложенной комплексной модели диктора позволило повысить вероятность идентификации более чем на 3% по сравнению с моделями, не учитывающими разделение по широким фонетическим классам.

Для обеспечения идентификации по голосу в режиме реального времени выполнено структурирование множества моделей диктора в виде иерархической базы данных. Оно представляет собой дерево с итеративным обобщением моделей. Листом дерева является модель речи диктора, узлом — фоновая модель речи некоторой группы дикторов, а корнем — обобщенная или универсальная фоновая модель. Были разработаны соответствующие методы манипуляции записями: процедуры добавления, удаления и редактирования записей. Для предотвращения неверного структурирования, возможного после добавления нескольких записей в базу данных, и с целью повышения эффективности идентификации реализована процедура оптимизации структуры базы данных.

Как показали численные исследования, процедура оптимизации может не только увеличить эффективность идентификации за счет корректировки состава кластеров признакового пространства, но и уменьшить глубину дерева, а следовательно, увеличить скорость поиска по нему.

SUMMARY

Yermolenko T.V., Klymenko M.S **Usage of Speech Signal Segmentation for the Construction of Complex Model in the Speaker Identification System.**

The article proposes an approach to creation of a complex speaker model based on Gaussian mixture. The elements of Gaussian mixture are formed by preliminary segmented speech signal, where each segment matches to certain broad phonetic class.

Efficiency of speaker identification system is evaluated in terms of the accuracy and speed of identification. It is shown that efficiency of speaker identification system depends on structure of speaker model, selected acoustic characteristics and classifiers, an approach to structuring the set of speaker models.

The mel-frequency cepstral coefficients are chosen as a feature set for describing of acoustic characteristics and Gaussian mixture is choosen as classifier.

The authors offer to use a set of Gaussian mixtures as a structure of speaker model. Each mixture of the set describes a particular broad phonetic class of speaker voice. The construction of models is performed using the subsystem of preliminary speaker-independent segmentation and classification of the speech signal.

Evaluation of speaker identification system shows that proposed complex speaker model allows to increase the probability of identifying by more than 3%.

We structure the set of speaker models as hierarchical database in order to provide real-time speaker identification. This database is organized as a tree with an iterative model generalization. A leaf of the tree stores a speaker model. Each internal node contains a background model of it's child nodes. The root node contains universal background model. Following manipulation methods have been developed: insert, delete and edit of records. After adding some records to the database it is possible that database structure become not optimal. We use an optimize procedure of the database structure to prevent this and improve the efficiency of identification.

As shown by evaluations, the optimize procedure increases the effectiveness of speaker identification by adjusting the composition of clusters and also reduces the depth of the tree, and thus increases the speed of searching on it.