УДК 004.934.8’1

# СИСТЕМА ИДЕНТИФИКАЦИИ ДИКТОРОВ НА ОСНОВЕ ОБЪЕДИНЕНИЯ ПРИЗНАКОВ, ВЕКТОРНОГО КВАНТОВАНИЯ И НОРМАЛИЗАЦИИ РАССТОЯНИЙ

**Первушин Е.А.**

*ГОУ ВПО «Омский государственный университет им. Ф.М. Достоевского», Омск, e-mail: pervushin_evgen@mail.ru*

Статья посвящена разработке системы идентификации дикторов. Во введении дается общее представление об основных элементах системы, а также подчеркивается актуальность задачи повышения точности распознавания. Далее определяются методы извлечения признаков, которые используются в данной работе. Предлагается к исследованию подход, при котором признаки речевого сигнала, полученные от разных алгоритмов, комбинируются для создания более точного представления. При этом происходит объединение множеств векторов без изменения размерностей самих векторов признаков. Далее описывается используемая модификация алгоритма векторного квантования, основанная на алгоритме K-средних. Разработанная модификация позволяет кластеризовать векторы переменной длины. Также в традиционную систему, основанную на вычислении расстояний, вводится дополнительный этап, заключающийся в нормализации расстояний, получаемых различными звуковыми фрагментами. Описывается серия экспериментов, моделирующих различные условия использования системы. Оценка точности производится с помощью взвешенной суммы процентов верных идентификаций в экспериментах с различными условиями. Результаты экспериментов показывают, что за счет объединения признаков и выбора способа нормирования степеней подобия точность идентификации повышена с 81,53 до 89,8 %.

**Ключевые слова: идентификация дикторов, распознавание дикторов, извлечение признаков, векторное квантование, нормирование расстояния**

# SPEAKER IDENTIFICATION SYSTEM BASED ON FEATURE JOINING, VECTOR QUANTIZATION AND DISTANCE NORMALIZATION

**Pervushin E.A.**

*Omsk state university of F.M. Dostoevsky, Omsk, e-mail: pervushin_evgen@mail.ru*

This paper presents a text-independent speaker identification system. Recognition scheme is based on vector quantization. In this article we provide modification of this method that allows for clustering of variable length vectors. Also, we propose two approaches. In first approach, a number of algorithms is used for feature extraction. The sets of various features present a single model. This feature joining is intended for more informative presentation of signal. In second approach, methods of distance normalization are introduced. Methods transform distances into matching scores using distances to all models. Distances from various speech segments are normalized in such way. Finally, we describe closed-set speaker identification experiments. The experiments conducted with 50 speakers. Weighted sum of identification rate in various conditions is used as performance measure. We show how new approaches contribute to the overall system performance. The achieved performance is 89,8 % while baseline system performance is 81,53 %.

**Keywords: speaker identification, speaker recognition, feature extraction, vector quantization, distance normalization**

Задача идентификации дикторов заключается в определении по образцу записи голоса, кому из ранее зарегистрированных пользователей принадлежит данный образец. Данная задача решается с помощью совокупности методов, в работе которых могут быть выделены следующие этапы: обработка сигнала с целью выделения векторов признаков, создание модели диктора и определения метода сравнения между извлекаемыми признаками и моделями дикторов, а также методы принятия решения на основе полученных сравнений.

Основной целью исследований в области распознавания дикторов является создание алгоритмов, повышающих точность работы систем, сохраняя при этом приемлемые показатели по вычислительной трудоемкости. В данной работе исследуется несколько способов. Во-первых, исследуется возможность повышения точности за счет одновременного использования векторов признаков, полученных от разных алгоритмов обработки сигналов. Во-вторых, исследуются способы нормирования расстояний, используемых для принятия решения об идентификации.

## Извлечение признаков

На этапе извлечения признаков речевой сигнал сегментируется на короткие участки и на каждом участке вычисляется набор признаков. В области распознавания дикторов наибольшую популярность приобрели кепстральные методы извлечения признаков: на основе линейного предсказания (LPCC) и мэл-частотные (MFCC) (см., например, [2, 5]). В проведенных экспериментах анализ проводился на участках размером 16 мс, перекрытие между окнами 8 мс. Для получения коэффициентов LPCC вычислялось по 20 коэффициентов линейного

предсказания, из которых генерировалось по 30 кепстральных коэффициентов. Также использовалось 32 коэффициента MFCC, вычисленных по значениям логарифмов энергий 48-ми подполос диапазона 0–5000 Гц.

Помимо описанных методов, в данной работе применялся также метод, основанный на выделении в речевом сигнале участков, соответствующих периодам основного тона. Такой подход был исследован в [1] и показал результаты, сравнимые по точности с коэффициентами MFCC.

В данной работе исследуется подход, при котором признаки речевого сигнала, полученные от разных алгоритмов, комбинируются для создания более точного представления. Для этого объединяются последовательности извлекаемых векторов. Размерность векторов признаков каждого алгоритма не меняется.

### Создание модели диктора

В данной работе для получения модели диктора использовался метод векторного квантования с использованием алгоритма K-средних [4, 6] для кластеризации данных.

Пусть $\mathbf{v}_1, \ldots, \mathbf{v}_L$ – входная последовательность векторов. Начальные значения средних инициализируем векторами из исходной последовательности с индексами $\left\lfloor \frac{L}{K}(\frac{1}{2}+i) \right\rfloor$, $i = 0, \ldots, K-1$. В проведенных экспериментах было использовано значение $K = 96$ кластеров. Нахождение ближайших кластеров определим на основе евклидова расстояния в случае векторов фиксированной размерности. Использование векторов произвольной длины (как в случае с описанным выше методом на основе кадров периода основного тона) требует некоторой модификации алгоритма. Во-первых, при вычислении расстояния между двумя такими векторами вектор с большим количеством координат обрезается, а сумма усредняется по количеству слагаемых. Во-вторых, для вычисления среднего определим вычисление суммы следующим образом. Пусть $\mathbf{S}$ – вычисленная до данной итерации сумма n векторов (возможно, пока содержащая только один вектор), $L_S$ – количество координат вектора $\mathbf{S}$, $\mathbf{v}$ – вектор, состоящий из $L_v$ координат. Тогда вектор суммы $\mathbf{S} + \mathbf{v}$ будет содержать $L = \max(L_\mathbf{S}, L_\mathbf{v})$ координат, вычисленных по формуле

$$(\mathbf{S}+\mathbf{v})_i = \begin{cases} \mathbf{S}_i + \mathbf{v}_i, & 1 \le i \le \min(L_\mathbf{S}, L_\mathbf{v}), \\ \mathbf{S}_i \dfrac{n+1}{n}, & L_\mathbf{v} < i \le L, \\ \mathbf{v}_i(n+1), & L_\mathbf{S} < i \le L. \end{cases}$$

В случае, если речевой сигнал обрабатывается несколькими алгоритмами извлечения признаков, множество кластеров (кодовая книга) строится отдельно для каждой извлекаемой последовательности.

### Классификация

Рассмотрим теперь процесс идентификации. Образец речевого сигнала обрабатывается и представляется с помощью последовательности векторов $\mathbf{v}_1, \ldots, \mathbf{v}_L$. От каждого вектора $\mathbf{v}_i$ вычисляются кратчайшие расстояния до шаблонов каждого диктора. Для этого используем евклидово расстояние, усредненное по количеству элементов. Обозначим через $d_{ij}$ – расстояние от вектора $i$ до шаблона $j$. Традиционный подход к классификации на основе векторного квантования или метода ближайшего соседа заключается в вычислении среднего по векторам расстояния до шаблонов [3].

Расстояния до конкретного шаблона, полученные от векторов, соответствующих различным звуковым фрагментам, могут существенно различаться. Заметим также, что при объединении векторов признаков из разных признаковых пространств процедура нормализации расстояний становится необходимой. Поэтому предлагается осуществлять процедуру нормализации расстояний. Нормализацию на уровне принятия решений можно сравнить с нормализацией, осуществляемой в задаче верификации, в которой вычисляются отношения степеней подобия предъявленного образца и заявленной идентичности к степени подобия с некоторым множеством референтных пользователей, называемым также когортой. В задаче идентификации нет необходимости отдельно хранить когортные модели, для фиксированного расстояния $d_{ij}$ в качестве референтных выступают расстояния $d_{ik}$, $k \neq j$.

При использовании отношения между расстояниями удобно перейти к термину «степень подобия». Степень подобия тем выше, чем короче расстояние. Рассмотрим несколько способов вычисления степеней подобия. Каждый из них можно рассматривать как вектор-функцию, при данных значениях расстояний $(d_{i,1}, ..., d_{i,N})$ вычисляющую степени подобия $(s_{i,1}, ..., s_{i,N})$. Полученные для исходных векторов степени подобия затем суммируются для принятия итогового решения

$$C = \arg\max_{1 \le j \le N} \sum_{i=1}^{L} s_{i,j}.$$

Пусть найдены кратчайшие расстояния $(d_{i,1}, ..., d_{i,N})$ от вектора $\mathbf{v}_i$ до хранимых шаблонов. Для дальнейшего использования упорядочим расстояния по неубыванию:

$(d'_{i,1},\ldots,d'_{i,N})$, $d'_{i,j} \le d'_{i,j+1}$. Функцию вычисления степеней подобия зададим следующим образом

$$s_{i,j} = \begin{cases} \dfrac{d_c - d_{i,j}}{d_c - d'_{i,1}}, & d_{i,j} < d_c, \\ 0, & d_{i,j} \ge d_c. \end{cases} \qquad (1)$$

Здесь $d_c > d'_{i,1}$ – некоторое расстояние, используемое для нормирования. В проведенных экспериментах расстояние $d_c$ выбиралось как элемент (координата) с определенным индексом или среднее по нескольким первым элементам вектора упорядоченных расстояний.

Следующий способ иногда называют схемой голосования. Пусть для вектора $\mathbf{v}_i$ идентифицируемой последовательности найдено $k$ ближайших векторов среди хранимых шаблонов, $k_{ij}$ – количество векторов среди найденных, принадлежащих шаблону $j$. Тогда положим $s_{ij} = k_{ij}/k$. Использованное в экспериментах значение $k$ равно единице.

Проблема выбора наиболее подходящего метода и его параметров может быть решена при достаточном количестве доступных для обучения данных с помощью метода кросс-валидации. Настройка параметров на этапе принятия решения упрощается тем, что данный этап является заключительным в процессе идентификации.

### Эксперименты и результаты

Эксперимент по оценке точности работы был проведен на речевой базе данных, содержащей образцы речи, записанные в офисных условиях. Частота дискретизации записей составляет 16 кГц. База содержит мужские и женские голоса. Использовались записи пятидесяти дикторов. Каждый из дикторов записал по две сессии, интервал между которыми составляет не менее суток. В каждую сессию была сделана запись на два разных микрофона (электретный и динамический).

Для того чтобы смоделировать различные условия применения системы, было поставлено несколько экспериментов, результаты которых для целей сравнения и определения наиболее результативных параметров были объединены во взвешенную сумму.

В первой части экспериментов для обучения моделей и для попыток идентификации использовалась короткая фиксированная фраза, одинаковая для всех дикторов. Продолжительность произнесения фразы 3–5 с.

Во второй части экспериментов в качестве материала для произнесения каждому пользователю предоставлялись различные для каждой сессии тексты. Для обучения моделей использовалось по 40 с речи. Для попыток идентификации первые 40 с записи второй сессии разбивались на четыре десятисекундных сегмента. Обе записи получены с использованием одного микрофона.

Последняя часть экспериментов повторяет схему второй части за исключением того, что используемый для регистрации материал записан с использованием другого микрофона.

В каждом эксперименте для регистрации пользователей в системе использовались записи первой сессии, записи второй сессии использовались для проведения тестовых оценок. Для исследования влияния количества зарегистрированных пользователей эксперименты проводились по группам с количеством пользователей, равным 10, 25 и 50 для каждой из описанных частей экспериментов. Для объединения результатов было решено назначать весовые коэффициенты в зависимости от количества зарегистрированных пользователей: $w_1 = 1$ в экспериментах по идентификации среди десяти пользователей, $w_2 = 2$ – среди двадцати пяти пользователей и $w_3 = 3$ – среди пятидесяти пользователей. Используемые значения весовых коэффициентов нормируются так, что их сумма равна единице.

Результаты экспериментов приведены в таблице. Эксперименты проведены для методов извлечения признаков LPCC (L), MFCC (M), кадров основного тона (F), а также их попарных объединений. Указаны взвешенные по всем экспериментам проценты верных идентификаций, а также результаты отдельных экспериментов для пятидесяти дикторов из второй и третьей частей экспериментов. Приведены несколько функций вычисления степеней подобия:

- $F_1$ – традиционный способ, расстояния суммируются;
- $F_2$ – схема голосования;
- $F_3$ – вычисляется степень подобия по формуле (1), в которой положено $d_c = d'_{\lfloor 0{,}2N \rfloor}$.

В качестве исходной точки для сравнения следует рассматривать результаты, соответствующие признакам LPCC или MFCC с использованием функции $F_1$. Результаты экспериментов показывают, что точность идентификации может быть повышена как за счет объединения признаков, так и за счет выбора способа нормирования степеней подобия.

Результаты экспериментов по идентификации дикторов

| Признаки | Взвешенный процент по всем экспериментам | | | Отдельные эксперименты | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | совпадающие условия, 50 дикторов | | | несовпадающие условия, 50 дикторов | | |
| | $F_1$ | $F_2$ | $F_3$ | $F_1$ | $F_2$ | $F_3$ | $F_1$ | $F_2$ | $F_3$ |
| L | 80,69 | 83,72 | 84,08 | 90,5 | 89,5 | 92,5 | 56,0 | 57,0 | 61,5 |
| M | 81,53 | 86,86 | 87,28 | **94,0** | 92,5 | 93,0 | 56,0 | 62,5 | 69,5 |
| F | 76,69 | 73,75 | 75,94 | 88,0 | 83,0 | 84,5 | 51,0 | 38,0 | 44,0 |
| L + M | **84,33** | **89,33** | 89,44 | **94,0** | **95,5** | 94,5 | **58,5** | **65,5** | **73,5** |
| L + F | 76,36 | 87,08 | **89,8** | 88,0 | 93,0 | **96,0** | 51,0 | 62,0 | 72,0 |
| M + F | 76,36 | 86,25 | 88,58 | 88,0 | 90,5 | 91,5 | 51,0 | 61,0 | 71,5 |

## Заключение

Проведенные в данной работе исследования выявили несколько способов повышения точности систем распознавания. Первый из них заключается в объединении последовательностей векторов признаков, полученных от разных алгоритмов извлечения признаков.

В процессе работы над системой распознавания была создана модификация алгоритма K-средних, позволяющая кластеризовать векторы переменной длины.

Рассмотренные способы нормализации степеней подобия, получаемых векторами различных участков речевого сигнала, также позволяют увеличить точность идентификации. Рассмотренная функция кусочно-линейна, в качестве дальнейшего направления работы могут быть проведены исследования, использующие более широкий класс функций.

### Список литературы

1. Первушин Е.А. Извлечение признаков во временной области в системе распознавания дикторов // Информационные технологии и автоматизация управления: материалы III научно-практической конференции. – 2011. – С. 265–267.

2. Kinnunen T., Spectral features for automatic text-independent speaker recognition. – Licentiate thesis, Department of Computer Science, University of Joensuu, Joensuu, Finland. – 2003.

3. MacQueen J. Some methods for classification and analysis of multivariate observations // In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability (University of California, 1967). – Vol. I. – P. 281–297.

4. Gill M.K., Kaur R., Kaur J. Vector Quantization based Speaker Identification // International Journal of Computer Applications. – 2010. – Vol.4, № 2.

5. Kinnunen T. Spectral features for automatic text-independent speaker recognition. Licentiate thesis, Department of Computer Science, University of Joensuu, Joensuu, Finland. – 2003.

6. MacQueen J. Some methods for classification and analysis of multivariate observations // Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability (University of California, 1967). – Vol. I. – P. 281–297.

**Рецензенты:**

Гуц А.К., д.ф.-м.н., профессор, декан факультета компьютерных наук Омского государственного университета им. Ф.М. Достоевского, г. Омск;

Горлов С.И., д.ф.-м.н., профессор, ректор Нижневартовского государственного гуманитарного университета, г. Нижневартовск;

Захарченко В.Д., д.т.н., профессор, профессор кафедры радиофизики Волгоградского государственного университета, г. Волгоград.

Работа поступила в редакцию 11.07.2011.