Доклады БГУИР
2003 *октябрь–декабрь* *№ 4*

УДК 681.3

МОДЕЛИ ГАУССОВЫХ СМЕСЕЙ ДЛЯ ВЕРИФИКАЦИИ ДИКТОРА ПО ПРОИЗВОЛЬНОЙ РЕЧИ

Р.Х. САДЫХОВ, В.В. РАКУШ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 16 июня 2003

Отдельные гауссовы компоненты хорошо моделируют характеристики голоса диктора, необходимые для идентификации по произвольной речи. В статье представлена математическая модель гауссовых смесей для задач распознавания диктора и предложена интерпретация компонентов модели. Приводится алгоритм обучения моделей и предложена реализация системы распознавания. Результаты экспериментов демонстрируют эффективность моделей гауссовых смесей порядка 94 % для распознавания голоса диктора по произвольной речи. В качестве базы данных для проверки эффективности алгоритмов была использована речь, записанная на телефонном канале.

Ключевые слова: распознование диктора, модели гауссовых смесей, векторное квантование.

Введение

В статье описана система верификации диктора по голосу и представлены исследования использования моделей гауссовых смесей для задачи распознавания диктора по голосу. Хотя существуют другие методы распознавания диктора, основанные на нейронных сетях [1, 2] или статистике, вычисленной на продолжительном интервале [3, 4, 15], модели гауссовых смесей хорошо себя зарекомендовали в качестве стохастической модели для построения систем распознавания [8]. Во-первых, модели очень удобны для моделирования не только статистических характеристик голоса диктора, но и окружающей среды, канала звукозаписи [9, 14]. Во-вторых, смеси гауссовых моделей представляют собой удобный способ представления и интерпретации акустических событий речевого сигнала.

В статье представлен новый алгоритм оценки параметров моделей гауссовых смесей на основе алгоритма *k*-средних. Представлено теоретическое доказательство сходимости нового алгоритма. Также приведено экспериментальное сравнение эффективности распознавания при обучении моделей общепринятым алгоритмом оценки максимизации и алгоритмом *k*-средних.

Одной из проблем при построении систем распознавания на основе модели гауссовых смесей является проблема инициализации параметров модели. В статье предлагается новый алгоритм инициализации, основанный на алгоритме построения кодовой книги векторного квантователя из множества векторов обучения.

Кроме алгоритмов обучения системы и экспериментов по построению систем распознавания голоса диктора на основе модели гауссовых смесей в статье приводятся эксперименты по выбору вектора параметров. Наиболее часто используемый вектор параметров состоит из кепстральных коэффициентов, но для повышения устойчивости системы распознавания к аддитивному шуму часто в качестве параметров вектора используют первую производную кепстральных коэффициентов по времени, нормализацию среднего значения

кепстральных коэффициентов [10]. Эффективность распознавания системы идентификации сильно зависит от выбора параметров, характеристик канала записи речи, эмоционального и физического состояния диктора. Не существует теоретического способа определения оптимальной структуры системы идентификации, и поэтому настройку системы производят экспериментально. В статье представлены результаты экспериментов зависимости эффективности распознавания от типа используемых параметров, их числа и размерности вектора параметров. Приведены экспериментальные исследования зависимости эффективности распознавания от числа гауссовых смесей модели.

Экспериментальная проверка разработанных алгоритмов проведена на основе систем верификации диктора по голосу. Эффективность системы верификации диктора по голосу может быть оценена при помощи графика оперативной характеристики надежности (ОХН), введенной в исследования по распознаванию диктора из области исследований психофизики. График ОХН получается путем откладывния на двух осях координат двух значений вероятностей: вероятности распознавания своего диктора как чужого и вероятности распознавания чужого диктора как своего, при изменении порога распознавания [8].

Смеси гауссовых моделей для распознавания диктора

Модель гауссовых смесей представляет собой взвешенную сумму M, компонент и может быть записана выражением

$$p(\bar{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\bar{x}), \qquad (1)$$

где $\bar{x}$ — это D-мерный вектор случайных величин; $b_i(\bar{x}), i = 1,...,M$, — функции плотности распределения составляющих модели и $p_i, i = 1,...,M$, — веса компонентов модели. Каждый компонент является D-мерной гауссовой функцией распределения вида

$$b_i(\bar{x}) = \frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{x}-\bar{\mu}_i)^T \sum_i^{-1} (\bar{x}-\bar{\mu}_i)\right\}, \qquad (2)$$

где $\bar{\mu}_i$ — вектор математического ожидания и $\sum_i$ — ковариационная матрица. Веса смеси удовлетворяют выражению $\sum_{i=1}^{M} p_i = 1$.

Полностью модель гауссовой смеси определяется векторами математического ожидания, ковариационными матрицами и весами смесей для каждого компонента модели. Эти параметры все вместе записываются в виде

$$\lambda = \left\{p_i, \bar{\mu}_i, \sum_i\right\}, i = 1,...,M. \qquad (3)$$

В задаче распознавания диктора каждый диктор представляется моделью гауссовых смесей и ставится в соответствие со своей моделью λ. Модель гауссовой смеси может иметь несколько различных форм в зависимости от вида ковариационной матрицы. Модель может иметь одну ковариационную матрицу для каждого компонента модели, как определено в (3), одну ковариационную матрицу для всех гауссовых компонент в модели или одну ковариационную матрицу, используемую всеми дикторами во всех моделях. Ковариационная матрица также может быть полной или диагональной. В экспериментах используется диагональная ковариационная матрица для каждого отдельного компонента модели.

Существует две причины использования моделей гауссовых смесей для идентификации диктора. Первая причина является интуитивным предположением того, что отдельные компоненты модели могут моделировать некоторое множество акустических признаков/событий. Можно предположить, что акустическое пространство голоса диктора может быть характеризовано множеством классов, представляющих некоторые фонетические

события/звуки как гласные, фрикативные и т.д. Эти акустические классы отражают некоторые общие, но особенные для каждого диктора конфигурации голосового тракта, и поэтому они эффективны для идентификации диктора. Спектр акустического класса может быть представлен вектором математического ожидания, а изменение среднего спектра может быть представлено ковариационной матрицей. Поскольку звуки речи, используемой для обучения или распознавания, не имеют пометок, то фонетические события "зашифрованы" в классах акустического пространства. Предполагая, что векторы признаков не зависимы друг от друга, плотность наблюдения векторов, образующих эти классы, можно считать смесью гауссовых распределений.

Второй причиной использования моделей гауссовых смесей для идентификации диктора является эмпирическое наблюдение, что линейная комбинация гауссовых распределений может представлять большое число классов акустических признаков. Одна из сильных сторон смеси гауссовых моделей та, что эти модели могут очень точно аппроксимировать произвольные распределения. Классическая модель представления диктора при помощи одного гауссова распределения описывается при помощи позиции распределения (вектора математических ожиданий) и формой распределения (ковариационной матрицей). Модель векторного квантования представляет диктора при помощи дискретного множества кластеров кодовой книги. В некотором смысле модель гауссовых распределений представляет собой гибрид между этими двумя моделями (векторного квантования и гауссова распределения), так как использует дискретное множество гауссовых функций. Каждая функция имеет собственную величину вектора математических ожиданий и ковариационную матрицу.

Поскольку гауссовы смеси моделируют функцию плотности вероятности вместе, то нет необходимости использовать полные ковариационные матрицы, даже если параметры вектора не являются полностью независимыми друг от друга. Линейная комбинация диагональных ковариационных матриц способна моделировать корреляцию между элементами вектора наблюдений. Эффект использования множества M ковариационных матриц может быть достигнут путем увеличения числа гауссовых компонент, использующих диагональные ковариационные матрицы.

Реализация алгоритмов оценки параметров моделей. Алгоритм оценки максимизации

Цель алгоритма оценки параметров модели — это при заданном обучающем высказывании диктора оценить параметры модели λ, которые наилучшим образом соответствуют распределению векторов признаков обучающего высказывания. Существует несколько способов оценки параметров модели, но наиболее популярным и широко используемым является метод оценки максимального правдоподобия.

Цель оценки максимального правдоподобия — найти параметры модели, которые максимизируют правдоподобие этой модели, при заданных обучающих данных. Для последовательности обучающих векторов $X = \{\bar{x}_1,...,\bar{x}_T\}$ правдоподобие модели гауссовых смесей может быть записано в виде

$$p(X|\lambda) = \prod_{t=1}^{T} p(\bar{x}_t|\lambda). \qquad (4)$$

К сожалению, это выражение представляет нелинейную функцию от параметров λ, и ее непосредственное вычисление невозможно. Поэтому оценки параметров могут быть получены итерационно при помощи алгоритма оценки-максимизации [5].

Алгоритм оценки-максимизации начинается с оценки начальной модели λ, и затем вычисляются новые параметры модели $\bar{\lambda}$, такие, что $p(X|\bar{\lambda}) \geq p(X|\lambda)$. Новая модель затем становится начальной моделью для следующей итерации, и процесс переоценки параметров повторяется, пока не будет достигнут некоторый порог сходимости. Этот способ используется для оценки параметров скрытых марковских моделей при помощи алгоритма Баума-Велча [6].

На каждой итерации алгоритма оценки-максимизации используются следующие формулы переоценки параметров:

$$\overline{p}_i' = \frac{1}{T}\sum_{t=1}^{T} p\left(i\middle|\overline{x}_t,\lambda\right), \tag{5}$$

$$\overline{\mu}_i' = \frac{\sum_{t=1}^{T} p\left(i\middle|\overline{x}_t,\lambda\right)\overline{x}_t}{\sum_{t=1}^{T} p\left(i\middle|\overline{x}_t,\lambda\right)}, \tag{6}$$

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T}\left(\overline{x}_i - \overline{\mu}_i'\right)^2 p\left(i\middle|\overline{x}_t,\lambda\right)}{\sum_{t=1}^{T} p\left(i\middle|\overline{x}_t,\lambda\right)}, \tag{7}$$

где σ_i^2, x_t, μ_i — произвольные элементы векторов $\overline{\sigma}_i^2, \overline{x}_t, \overline{\mu}_i$ соответственно.

Апостериорная вероятность i-го акустического класса задается выражением

$$p\left(i\middle|\overline{x}_t,\lambda\right) = \frac{p_i b_i\left(\overline{x}_t\right)}{\sum_{k=1}^{M} p_k b_k\left(\overline{x}_t\right)}. \tag{8}$$

Существуют две проблемы при обучении смесей гауссовых моделей — это выбор числа компонентов модели и инициализация параметров моделей. Не существует теоретического решения этих задач.

Для идентификации диктора по голосу группа дикторов $S = \{1,2,...,S\}$ представляется набором моделей гауссовых смесей $\lambda_1, \lambda_2, ..., \lambda_S$. Цель идентификации — найти модель диктора, которая имеет наибольшее значение апостериорной вероятности для заданного высказывания:

$$S = \arg\max_{1\le k\le S} Pr\left(\lambda_k\middle|X\right) = \arg\max_{1\le k\le S} \frac{p\left(X\middle|\lambda_k\right)Pr\left(\lambda_k\right)}{p\left(X\right)}. \tag{9}$$

Предполагая, что все дикторы одинаково вероятны (т.е. $\Pr\left(\lambda_k\right) = 1/S$) и замечая, что величина значения $p\left(X\right)$ одинакова для моделей всех дикторов, правило классификации диктора упрощается до вида

$$S = \arg\max_{1\le k\le S} p\left(X\middle|\lambda_k\right). \tag{10}$$

Используя логарифм и независимость между наблюдениями, система идентификации диктора вычисляет

$$S = \arg\max_{1\le k\le S} \sum_{t=1}^{T} \log p\left(\overline{x}_t\middle|\lambda_k\right), \tag{11}$$

где $p\left(\overline{x}_t\middle|\lambda_k\right)$ определена в (1).

Инициализация параметров модели

При обучении модели гауссовых смесей существует проблема инициализации параметров модели перед началом процесса обучения. Алгоритмы обучения не гарантируют нахождение глобального максимума в пространстве векторов обучения и поэтому результат

обучения системы существенно зависит от начальных значений параметров системы. В статье предлагается метод инициализации параметров модели на основе векторного квантования.

Как описано в [11], при векторном квантовании N-мерный вектор параметров наблюдения $\bar{x}$ отображается в N-мерный вектор $\bar{y}$. Этот процесс называется квантованием. Множество векторов $Y = \{\bar{y}_i, 1 \le i \le L\}$ называется кодовой книгой преобразования или просто кодовой книгой. L — размер кодовой книги, а $\bar{y}_i$ — множество кодовых векторов. Для построения такой кодовой книги N-мерное пространство случайного вектора $\bar{x}$ разделяется на L областей или ячеек $\{C_i, 1 \le i \le L\}$ и с каждой такой ячейкой C_i связывается вектор $\bar{y}_i$. При квантовании вектора $\bar{x}$ квантователь назначает кодовый вектор $\bar{y}_i$, если x попадает в область C_i.

Для построения кодовой книги чаще всего используется алгоритм K-средних [12], в котором используется среднеквадратичное отклонение в качестве меры искажения:

$$D_i = \frac{1}{M} \sum_{x \in C_i} d(\bar{x}, \bar{y}_i), \tag{12}$$

$$d(\bar{x}, \bar{y}_i) = \frac{1}{N} \sum_{k=1}^{N} |x_k - y_k|^2 . \tag{13}$$

Алгоритм K-средних разбивает пространство обучающих векторов X на L кластеров. Как замечено в [13], кластеры кодовой книги можно рассматривать как описание событий/звуков речи. Также отдельные гауссовы компоненты можно рассматривать как описание событий речи. На основании этого предлагается использовать кодовую книгу для инициализации моделей гауссовых смесей.

Каждый компонент модели гауссовых смесей можно представить математическим ожиданием $\bar{\mu}_i$ и ковариационной матрицей $\sum_i$, используя выражение (2). Предполагается, что число компонент в модели гауссовых смесей совпадает с размером кодовой книги, т.е. $M = L$. Тогда математическое ожидание $\bar{\mu}_i$ каждого компонента инициализируется значениями элементов кодового вектора $\bar{y}_i$. Ковариационная матрица $\sum_i$ вычисляется на основании векторов, принадлежащих i-му кластеру кодовой книги. Веса компонентов p_i модели инициализируются числом векторов, попавших в кластер C_i. Так как веса должны удовлетворять выражению $\sum_{i=1}^{M} p_i = 1$, то каждый вес нормализуется общим числом векторов $\bar{x}$, участвующих в обучении модели.

Структура системы верификации диктора по произвольной речи

На рис. 1 изображена структура системы верификации диктора по произвольной речи. Система состоит из 7 функциональных модулей. Модуль вычисления признаков осуществляет предварительную обработку речи, спектральный анализ и формирует вектор первичных признаков речевого сигнала. На вход этого блока также подаются значения параметров анализа. Параметрами первичного анализа являются: способ спектрального анализа (кратковременное преобразование Фурье или коэффициенты линейного предсказания), число кепстральных коэффициентов, использование Δ-параметров. В табл. 1 приведены результаты верификации для различных типов первичных параметров. Для повышения процента распознавания при обучении и распознавании речи по каналам с различными шумовыми и амплитудно-частотными характеристиками в систему включен блок нормализации параметров. Речевой сигнал, передаваемый по телефонным линиям, подвержен сильным линейным искажениям,

которые можно записать в виде выражения $T(z) = S(z)G(z)$, где $S(z)$ соответствует спектру неискаженной речи, $G(z)$ — спектральной характеристике телефонного канала, и $T(z)$ — результирующему речевому сигналу. Прологарифмировав обе части этого выражения, получим

$$\log T(z) = \log S(z) + \log G(z). \quad (14)$$

Полагая, что речевой сигнал и спектральная характеристика канала могут быть представлены при помощи модели линейного предсказания с достаточно малой погрешностью, можно утверждать, что влияние канала на речевой сигнал приводит к аддитивной составляющей в LPC кепстре результирующего речевого сигнала $T(z)$. Далее, полагая, что математическое ожидание кепстральных коэффициентов неискаженной речи равно нулю, получаем, что оценка спектральной характеристики канала равна математическому ожиданию речи, искаженной каналом $T(z)$. Таким образом, компенсация влияния канала на речевой сигнал производится по формуле

$$c_{cms}(n) = c_{lp}(n) - E\left[c_{lp}(n)\right], \quad (15)$$

где $E[\bullet]$ — знак математического ожидания.

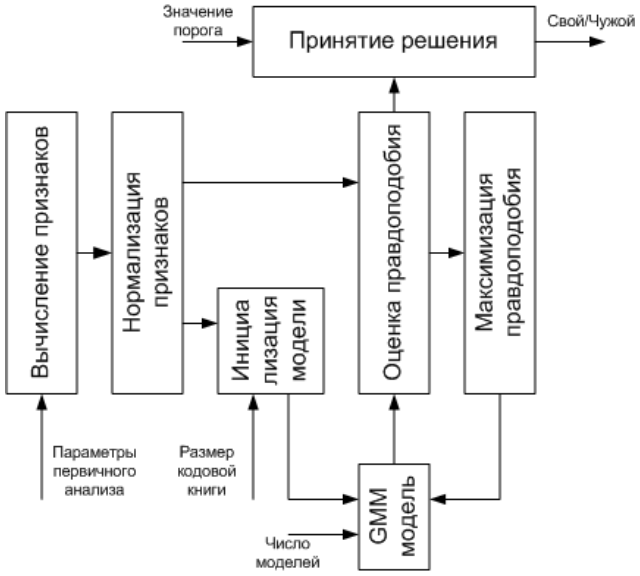

Рис. 1. Структура системы верификации диктора по произвольной речи

Блок инициализации параметров модели реализован на основании алгоритма векторного квантования, описанного раньше. Размер кодовой книги должен соответствовать числу смесей в модели. Блок GMM представляет собой блок памяти, в котором хранятся параметры модели. Для хранения параметров модели, состоящей из N смесей, требуется N+N×M×N×M×M ячеек памяти. M — это число элементов в векторе первичных признаков. Так при использовании 24 кепстральных коэффициентов и 24 Δ-кепстральных коэффициентов в качестве признаков требуется 307 328 байт. Предполагается, что число смесей 32 и для хранения числа с плавающей запятой используется 4 байта. Блок оценки правдоподобия вычисляет вероятность появления речевого высказывания для заданной модели смесей. Этот блок используется как при обучении модели, так и при тестировании высказывания. При обучении модели с помощью алгоритма оценки-максимизации в блоке максимизации вычисляются новые параметры модели. Процесс обучения требует около 10–15 итераций для нахождения эффективных значений парметров модели. В режиме тестирования высказывания оценка правдоподобия попадает в блок принятия решения, где сравнивается с пороговым значением. Если оценка больше значения порога, то диктор считается тем, за кого он себя выдает. Иначе считается, что диктор пытается обмануть систему.

Алгоритм оценки-максимизации является относительно ресурсоемким. Как было подсчитано, хранение 32 смесей модели требует примерно 300Кбайт памяти. В табл. 2. приведены оценки числа операций, необходимых для выполнения одной итерации алгоритма:

Таблица 1. **Эффективность верификации для различных типов первичных признаков**

Тип первичных признаков	Число параметров в векторе	EER, %
LPCC	12	15,4
LPCC+Δ	24	9,8
LPCC+ Δ+MCN	24	7,5
LPCC+ Δ+MCN	36	8,2
LPCC+ Δ+MCN	48	5,2

LPCC — кепстральные коэффициенты, полученные из коэффициентов, линейного предсказания; Δ — дельта параметры кепстральных коэффициентов; MCN — нормализация кепстральных коэффициентов при использовании выражение (15).

Таблица 2. **Вычислительная сложность алгоритма оценки-максимизации**

Тип операции	Чило оперций
Сложение/вычитание	(4×M×N+N)×T
Умножение	2×N×T+M×N+N+1
Вычисление экспоненты	N×T
Вычисление квадратного корня	N

Экспериментальная оценка

Был проведен ряд экспериментов, которые подтвердили эффективность предложенного метода инициализации модели гауссовых смесей и определили оптимальное число компонент модели. Также были проведены сравнения эффективности системы в зависимости от числа и типа параметров вектора первичных признаков.

Экспериментальная оценка эффективности системы производилась на базе высказывания, предложенного Oregon Institute of Science and Technology, Centre of Spoken Language Understanding. Экспериментальная выборка из базы голосов состояла из двух мужчин и двух женщин. Звуковой сигнал записывался с телефонной линии с частотой дискретизации 8 кГц и разрядностью АЦП 16 бит. Продолжительность обучающего высказывания составляла примерно 60 с. Продолжительность тестового высказывания составляла примерно 15 с. Для определения оптимального числа компонент сравнивались модели с числом компонент 2, 4, 8, 16, 24, 32. На рис. 2 изображена зависимость эффективности распознавания от числа компонент. Как видно из графика, рост вероятности распознавания существенно замедляется при значении числа компонент 32. Можно утверждать, что это значение является оптимальным для построения систем верификации. Это связано еще и с тем, что 60 с речи достаточно полно представляют все множество звуков речи, присущих диктору.

Для определения эффективности предложенного метода инициализации модели была выбрана модель с 32 компонентами. Две одинаковые системы верификации были проинициализированы различными способами: предложенным в статье и случайно выбранными из обучающей последовательности 32 векторами. Для каждой компоненты модели были сформированы множества ближайших векторов из обучающей последовательности. Далее проводилось обучение модели с использованием одинакового числа итераций. На рис. 3. изображены графики уровня ошибки для обоих способов инициализации. Из рисунка видно, что метод инициализации модели гауссовских смесей при помощи алгоритма векторного квантования улучшает эффективность верификации диктора на 0,5–1%.

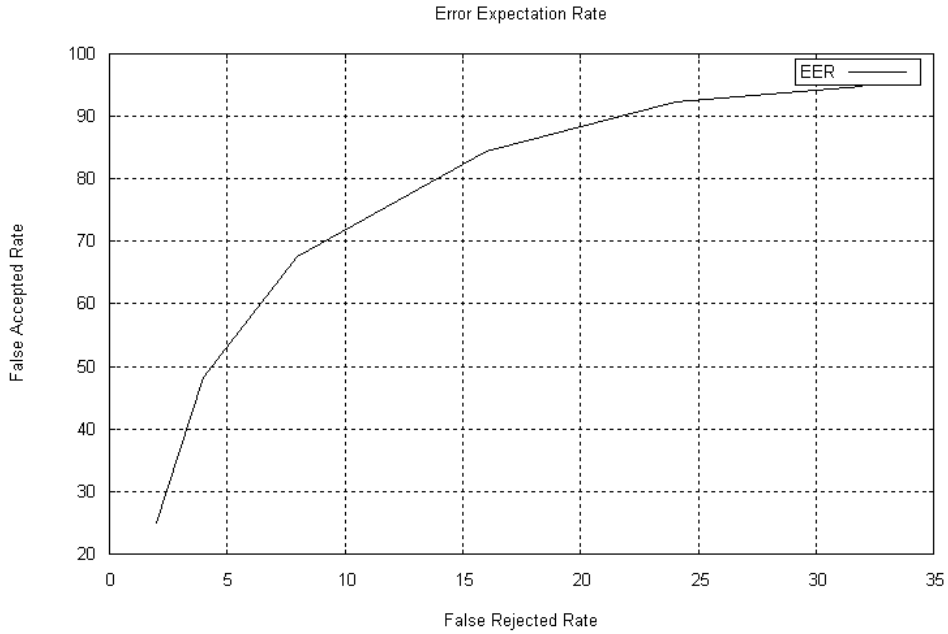

Рис. 2. Зависимость процента распознавания от числа смесей

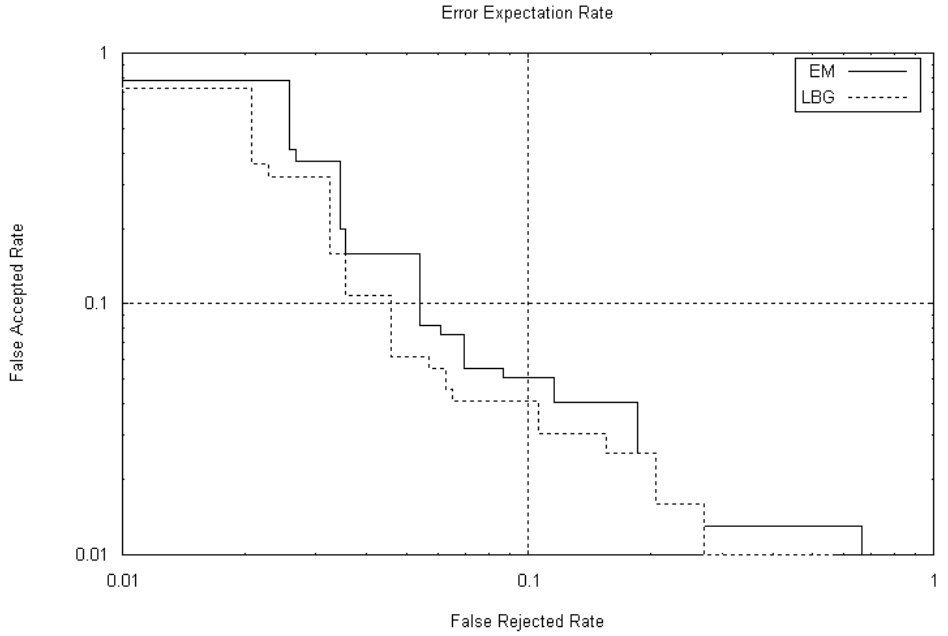

Рис. 3. Уровни ошибок для случайной инициализации и инициализации алгоритмом К-средних

Проводимые эксперименты показали значимость выбора начальных значений параметров моделей гауссовых смесей. При правильном выборе начальных значений существует большая вероятность того, что в процессе обучения будет достигнут глобальный максимум в пространстве параметров моделей или точка, близкая к нему. Для инициализации параметров был использован алгоритм векторного квантования. Кроме того, эксперименты показали необходимость создания быстрых алгоритмов обучения смесей гауссовых моделей.

Заключение

Система распознавания диктора, представленная в статье, имеет достаточно высокий процент распознавания и может быть использована во многих практических приложениях. Техника первичной обработки сигнала хорошо известна и легко реализуется на существующих процессорах. Данная система демонстрирует уровень распознавания, сравнимый, а иногда и превосходящий уровень других систем, описанных в литературе. Дальнейшие исследования в этой области могут быть направлены на разработку быстрых алгоритмов обучения моделей гауссовых смесей и на разработку алгоритмов адаптации к каналу звукозаписи и окружающей среды.

SPEAKER VERIFICATION ON ARBITRARY SPEECH USING GAUSSIAN MIXTURE MODELS

R.H. SADYKHOV, V.V. RAKUSH

Abstract

The separate Gaussian models are well known tool for modeling different stochastic processes. This paper describes mathematical background of the Gaussian mixture models for speaker verification tasks and provides models treatment. The models training algorithm and structure of the speaker recognition system were developed. The experimental results show performance about 94% for verification on arbitrary speech. Уexperiments has been done on the telephone speech database.

Литература

1. *Bennani Y., Fogelman Soulie F., Gallinari P.* // Confer. Proc. IEEE ICASSP. 1990. P. 265–268.
2. *Oglesby J., Mason J.S.* // Confer. Proc. IEEE ICASSP. 1990. P. 261–264.
3. *Markel J.D., Oshika B.T., Gray A.H.* // IEEE Trans. On Acoustics, Speech, and Signal Processing. 1977. Vol. 25. P. 330–337.
4. *Markel J.D., Davis S.B.* // IEEE Trans. On Acoustics, Speech, and Signal Processing. 1979. Vol. 27. P. 74–82.
5. *Dempster A., Laird N. Rubin D.* // J. Royal Stat. Soc. 1977. Vol. 39. P. 1–38.
6. *Baum L. et al.* // Ann. Math Stat. 1970. Vol. 41. P. 164–171.
7. *Reynolds D.A., Rose R.C.* // IEEE Trans. On Speech and Audio Proc. 1995. Vol. 3.
8. *Furui S.* // Digital Speech Processing, Synthesis and Recognition. Marcel Dekker, New York, 1989.
9. *Gopinath R.A.* // Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP). 1998. P. II.661–II.664.
10. *Mammone R.J., Zhang X., Ramachandran R.P.* // IEEE Signal Processing Magazine. Sept. 1996. P. 58–71.
11. *Makhoul J., Roucos S., Gish H.* // Proc. IEEE. 1985. Vol. 73. No. 11.
12. *MacQueen J.* // Proc. 5th Berkley Symp. on Math., Statist., and Prob. Berkley, CA: Univ. of California Press, 1967. P. 281–297.
13. Ракуш В.В., Садыхов Р.Х. // Тр. X науч.-техн. конф. "Новые технологии в машиностроении и вычислительной технике". Брест. 1992.
14. *Panda A., Bhattacharyya S., Shrikanthan T.* // Proc. of Symposium on Communication Systems, Networks and Digital Signal Processing. 2002. P. 383–386.
15. *Pelecanos J., Myers S., Sridharan S., Chandran V.* // Proc. of 15th Int. Conf. on Pattern Recognition. 2000. Vol. 3. P. 294–297.