электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

 # 06, июнь 2016

УДК 004.934

Выборочное распознавание фонем с помощью смеси гауссовых распределений

Ду Цзяньмин, *аспирант*
кафедра «Информационные системы и телекоммуникации»,
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана

Научный руководитель: Выхованец В.С., д.т.н., профессор
кафедра «Информационные системы и телекоммуникации»,
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
vykhovanets@bmstu.ru

Введение

Одним из крупных, наиболее востребованных и динамически развивающихся направлений является автоматическое распознавание речи, или преобразование речевого сигнала в текст. Можно выделить круг прикладных задач, решаемых в рамках автоматического распознавания речи. Это распознавание отдельных команд для управления устройствами и приложениями, распознавание отдельных фраз в системах массового обслуживания, поиск ключевых слов в системах мониторинга и распознавание слитной речи на большом словаре.

Несмотря на то, что распознавание отдельных команд, фраз и ключевых слов имеет эффективную практическую реализацию, задача достоверного распознавания слитной речи пока полностью не решена [**Ошибка! Источник ссылки не найден.**]. Отличительной особенностью применяемых для распознавания речи решений является необходимость выделения двух этапов: этапа обучения и, собственно, этапа распознавания [**Ошибка! Источник ссылки не найден.**].

Среди особенностей речи, приводящим к проблемам ее автоматического

распознавания, можно выделить две, более всего влияющие на конечный результат: высокую вариативность речи, обусловленную различием дикторов и изменчивостью окружающей акустической обстановки, а также высокую флективность (изменчивость) естественного языка. Следует отметить, что ненадежность распознавания является объективным фактором речевого способа коммуникации: при восприятии человек также не все распознает достоверно.

Настоящая статья посвящена технологии контекстного распознавания речи, которая основана на использовании грамматик естественных языков, описывающих отдельные предметные области устного дискурса [**Ошибка! Источник ссылки не найден.**]. В этом случае требования, предъявляемые к качеству сегментации речи и надежности распознавания фонем, могут быть ослаблены. По этой причине приемлемым являются более простые методы предварительной обработки речи, в частности, описанный в статье метод смеси гауссовых распределений GMM (англ. Gaussian mixture models) для выборочного распознавания фонем.

Суть подхода заключается в том, чтобы не распознать фонемы, а оценивать апостериорные вероятности появления фонем в каждом сегменте речевого сигнала. Временная последовательность списков фонем, отсортированных в порядке убывания апостериорной вероятности, используется далее для грамматического распознавания речевого фрагмента в целом. Базовыми точками в процессе такого распознавания являются фонемы, имеющие высокую апостериорную вероятность.

Смесь гауссовых распределений

Существует несколько оснований для использования смеси гауссовых распределений для распознавания фонем [**Ошибка! Источник ссылки не найден.**]. Во-первых, отдельные компоненты смеси могут моделировать акустические признаки речи, т.е. акустическое пространство речи может быть характеризовано множеством акустических классов. В этом случае предполагается, что векторы акустических признаков независимы, а плотность распределения векторов описывается смесью гауссовых распределений. Во-вторых, с помощью линейной комбинации гауссовых распределений может представлять большое число классов акустических признаков.

Плотность вероятности модели гауссовой смеси с порядком M описывается формулой [**Ошибка! Источник ссылки не найден.**]:

$$p(\mathbf{X}|\lambda) = \sum_{i=1}^{M} \omega_i b_i(\mathbf{X}), \quad (1)$$

где M – число компонентов, $\mathbf{X} = (x_1, x_2, \dots, x_D)$ – D-мерный вектор случайных величин (акустических признаков), λ – набор параметров распознаваемого класса (фонемы), ω_i – вес каждого распределения Гаусса (компонента смеси),

$$\sum_{i=1}^{M} \omega_i = 1,$$

где $b_i(\mathbf{X})$ – функции плотности вероятностей i-ой компоненты,

$$b_i(\mathbf{X}) = \frac{1}{(2\pi)^{D/2}|\mathbf{C}_i|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_i)^T \mathbf{C}_i^{-1}(\mathbf{X} - \boldsymbol{\mu}_i)\},$$

где $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{iD})$ – средний вектор i-ой компоненты, T – знак транспонирования, $\mathbf{C}_i$, $|\mathbf{C}_i|$ и $\mathbf{C}_i^{-1}$ – ковариационная матрица i-ой компоненты, ее определитель и обратная ей матрица.

Для распознания фонем могут использоваться диагональные ковариационные матрицы $\boldsymbol{\sigma}_i^2$. В этом случае модель гауссовой смеси для каждой фонемы определяется средними векторами $\boldsymbol{\mu}_i$, ковариационными матрицами $\boldsymbol{\sigma}_i^2$ и весами компонентов смеси ω_i,

$$\lambda = \{\omega_i,\ \boldsymbol{\mu}_i,\ \boldsymbol{\sigma}_i^2 \mid i = 1, 2, \dots, M\}.$$

Вектор акустических признаков

В качестве вектора признаков при распознавании фонем обычно используются мел-частотные кепстральные коэффициенты (англ. MFCC – Mel Frequency Cepstral Coefficients) [**Ошибка! Источник ссылки не найден.**]. Оцифрованный сигнал разбивается на блоки длительностью 25–30 мс. Пусть $y_0, y_1, \dots, y_{H-1}$ – отсчеты такого блока, где H – число отсчетов. К каждому блоку применяется некоторая весовая функция, например, окно Хэмминга

$$w_t = 0{,}54 - 0{,}46 \cdot \cos\left(2\pi \frac{t}{H-1}\right) \ (t = 0, 1, \dots, H-1),$$

и дискретное преобразование Фурье. Весовая функция используется для уменьшения искажений при преобразовании Фурье, вызванных конечностью выборки. Тогда коэффициенты дискретного преобразования Фурье взвешенного сигнала имеют вид:

$$s_k = \sum_{t=0}^{H-1} y_t w_t \exp(\frac{2\pi i}{H} kt) \ (k = 0, 1, \dots, H-1),$$

и соответствуют частотам $f_k = kF/H$, где F – частота дискретизации сигнала.

Полученное представление сигнала в частотной области разбивается на диапазоны

с помощью банка (гребенки) из V треугольных фильтров с весовыми коэффициентами u_{mk} $(k = 0,1, \dots, H-1;\ m = 0,1, \dots, V-1)$. Граничные частоты фильтров рассчитываются в мел-частотной области G по формуле

$$G(f) = 1127 \cdot \ln(1 + \frac{f}{700}).$$

После этого выполняется фильтрация квадратов модулей коэффициентов преобразования Фурье и логарифмирование полученного результата,

$$e_m = \ln(\sum_{k=0}^{H-1} |s_k|^2 u_{mk}) \ \ (m = 0, 1, \dots, V-1)\ ,$$

а сами MFCC коэффициенты x_k получаются путем дискретного косинусного преобразования,

$$x_k = \sum_{m=0}^{V-1} e_m \cos(\frac{\pi i(m+0{,}5)}{V}) \ \ (k = 0, 1, \dots, D-1).$$

Коэффициент x_0 является энергией сигнала и не используется. Количество коэффициентов D на практике выбирают от 12 до 30.

Обучение модели

В задаче распознавания фонем каждая фонема представляется своей гауссовой моделью со своим набором параметров λ. Пусть требуется распознать N фонем. Тогда все смеси гауссовых моделей будут задаваться набором параметров $\Lambda = \{\lambda_j \mid j = 1, 2, \dots, N\}$.

Распознавание фонемы φ сводится к тому, чтобы при заданной обучающей выборке оценить параметры модели λ_φ, которые наилучшим образом соответствуют распределению векторов признаков обучающей выборки.

Существует несколько способов оценки параметров модели, но наиболее широко используемым является метод оценки максимального правдоподобия, при котором ищутся параметры модели, которые при заданных обучающих данных максимизируют правдоподобие этой модели.

Для последовательности обучающих векторов $Q = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_S\}$, $\mathbf{X_t} = (x_{t1}, x_{t2}, \dots, x_{tD})$, правдоподобие каждой модели λ записывается так:

$$p(Q|\lambda) = \prod_{t=1}^{S} p(\mathbf{X}_t|\lambda). \qquad (2)$$

Из-за того, что $p(Q|\lambda)$ в формуле (2) является нелинейной функций параметров λ, оценки параметров λ получают итерационно при помощи алгоритма максимизации ожидания EM (англ. Expectation Maximization) [**Ошибка! Источник ссылки не найден.**].

Вычисления алгоритм EM начинает с начального значения λ, а затем вычисляются новые параметры модели λ' такие, что $p(Q|\lambda') \geq p(Q|\lambda)$. Новая модель затем становится начальной. Процесс переоценки параметров завершается, когда достигается некоторый порог сходимости.

На каждой итерации используются следующие формулы переоценки параметров

модели [**Ошибка! Источник ссылки не найден.**]:

$$\omega_i' = \frac{1}{S}\sum_{t=1}^{S} p_i(\mathbf{X}_t|\lambda),$$

$$\boldsymbol{\mu}_i' = \frac{\sum_{t=1}^{S} \mathbf{X}_t \cdot p_i(\mathbf{X}_t|\lambda)}{\sum_{t=1}^{S} p_i(\mathbf{X}_t|\lambda)},$$

$$\boldsymbol{\sigma}_i'^2 = \frac{\sum_{t=1}^{S} (\mathbf{X}_t - \boldsymbol{\mu}_i')^2 \cdot p_i(\mathbf{X}_t|\lambda)}{\sum_{t=1}^{S} p_i(\mathbf{X}_t|\lambda)},$$

где $p_i(\mathbf{X}_t|\lambda)$ – апостериорная вероятность i-ой компоненты,

$$p_i(\mathbf{X}_t|\lambda) = \frac{\omega_i' b_i(\mathbf{X}_t)}{\sum_{k=1}^{M} \omega_i' b_k(\mathbf{X}_t)} \quad (i = 1, 2, \dots, M).$$

При обучении смеси гауссовых моделей необходимо задать число компонентов модели M, весовые коэффициенты компонентов ω_i, а также априорные вероятности каждого компонента $p_i(\mathbf{X}_t|\lambda)$. Алгоритмы EM не гарантируют нахождение глобального максимума в пространстве обучающих векторов и поэтому результат обучения модели существенно зависит от начальных значений параметров.

Обычно используем случайные начальные значения. В проведенных экспериментах начальные средние вектора $\boldsymbol{\mu}_i$ выбрались случайно с равномерным распределением в пространстве обучающих векторов, начальные веса каждого распределения ω_i задавались одинаковыми, $\omega_i = 1/M$, а начальные ковариационные матрицы $\boldsymbol{\sigma}_i^2$ – определялись из обучающих данных Q,

$$\boldsymbol{\sigma}_i^2 = \left(\sigma_{ikk}^2 \middle| k = 1,2, \dots, D\right), \ \sigma_{ikk}^2 = \frac{1}{S}\sum_{t=1}^{S}(x_{tk} - \mu_{ik})^2.$$

На рис. 1 показан зависимость логарифма функции правдоподобия (2) от числа итераций при различных начальных условиях. Из рисунка видно, что алгоритм EM сходится достаточно быстро. Однако при различных начальных условиях получаются разные результаты, что объясняется наличием локальных экстремумов у функции правдоподобия. По этой причине при обучении используется множество начальных значений параметров и выбирается наилучший результат.

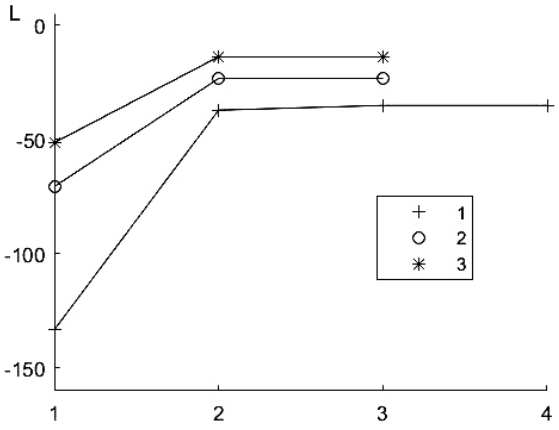

Рис. 1. Зависимость правдоподобия от числа итераций и начальных условий

В свою очередь выбор числа гауссовых компонентов M является более трудной задачей. Если значение M мало, то модель может быть грубой и приводить к ошибкам распознавания фонем. Если значение M велико, то число параметров модели будут значительным, что приводит к плохой сходимости алгоритма. В таблице показан процент распознанных фонем при различных M. Из таблицы видно, что имеется некоторое оптимальное число гауссовых компонентов, при которых число ошибок распознавания минимально. Экспериментально установлено, что для распознавания фонем достаточно использовать до 32 компонентов в каждой модели.

M	8	16	32	48
Доля, %	81,7	85,2	87,7	79,3

Распознавание фонем

Цель распознавания фонемы – найти из N моделей такую, которая имеет наибольшее значение апостериорной вероятности для заданного вектора акустических признаков $\mathbf{X}$,

$$\varphi = \arg \max_{1 \le j \le N} p(\lambda_j|\mathbf{X}) = \arg \max_{1 \le j \le N} \frac{p(\mathbf{X}|\lambda_j)p(\lambda_j)}{p(\mathbf{X})},$$

где $p(\lambda_j|\mathbf{X})$ – вероятность λ_j при условии $\mathbf{X}$, $p(\lambda_j)$ – априорная вероятность фонемы j. Так как вероятность $p(\mathbf{X})$ одинакова для всех моделей, правило распознавания упрощается,

$$\varphi = \arg \max_{1 \le j \le N} p(\mathbf{X}|\lambda_j)p(\lambda_j). \quad (3)$$

Структурная схема процесса обучения и распознавания фонем показана на рис. 2.

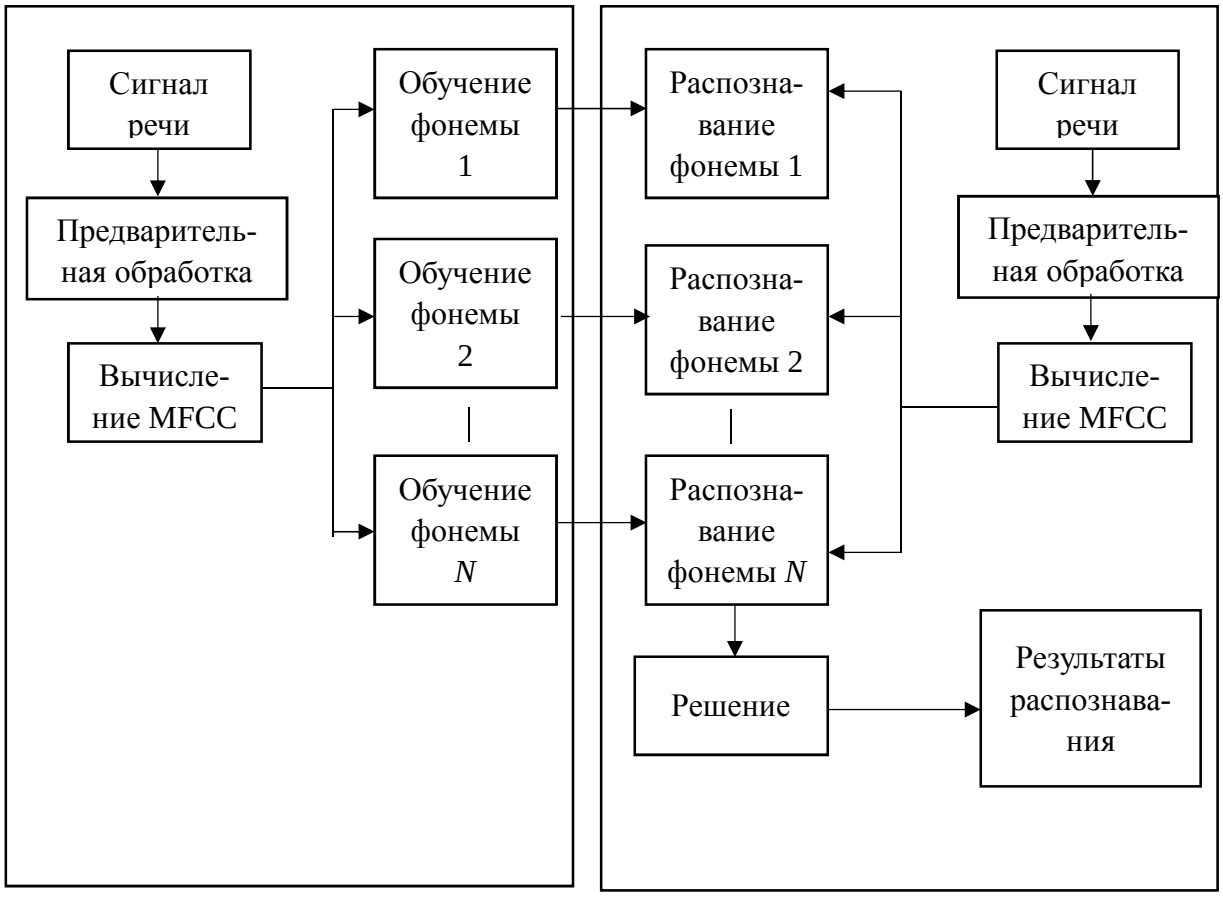

Рис. 2. Структурная схема распознавания фонем

На рис. 3 приведен пример результатов распознавания, откуда видно, что процесс распознавания фонем является непрерывным. Это позволяет избежать проблем, связанных с сегментацией речи.

Блок 76: м н
Блок 77: м н
Блок 78: ы и н м
Блок 79: И
Блок 80: И
Блок 81: И
Блок 82: И
Блок 83: И
Блок 84: ы и
Блок 85: и е ы
Блок 86: и ы

Рис. 3. Пример результатов распознавания фонем

Грамматическое распознавание речи

В связи с тем, что порядок слов в предложении русского языка не задан жестко, а может варьироваться без потери смысла предложения, грамматическое распознавание русской речи затруднено. В частности, затруднено использование статистических моделей языка на основе *N*-грамм, а также простых грамматик с жесткими грамматическими конструкциями [**Ошибка! Источник ссылки не найден.**].

В отличие от других известных подходов в грамматическом распознавании речи, в предлагаемом подходе планируется не ограничиваться простыми конструкциями языка (словами, словосочетаниями, устойчивыми оборотами), а использовать более сложные грамматические конструкции вплоть до сложных предложений.

Так как правильная интерпретация речи возможно только с учетом ситуативного контекста, в соответствии с жизненным опытом собеседников и их профессиональными навыками, то видится проблематичным создание и использование универсальных грамматик для распознавания речи в целом. Неизбежным компромиссом в этом случае является использование грамматик для небольших предметных областей, например, для диалога с роботом, где с целью повышения эффективности грамматического подхода значительно ограничены выразительные средства языкового общения.

Заключение

В настоящей статье предложен подход для распознавания слитной речи, основанный на использования обучаемой смеси гауссовых моделей для непрерывного определения апостериорных вероятностей не одной, а множества фонем в потоке речи, а также использования получаемой при этом временной последовательности множества фонем, отсортированных в порядке убывания апостериорных вероятностей, для грамматического распознавания речи с помощью грамматик, описывающих ограниченные области языкового общения.

Список литературы

[1]. Шелепов В.Ю. К проблеме пофонемного распознавания // Искусственный интеллект. 2005. № 4. С. 662-668.

[2]. Петровский А.А. Методы построения устройств распознавания речи на базе гибрида «нейронная сеть» – «скрытая Марковская модель» // Нейрокомпьютеры: разработка и применение. 2002. № 12. С. 26-36.

[3]. Выхованец В.С., Ду Цзяньмин, Назарова С.И. Контекстное распознавание слитной речи // Материалы VII-й Международной конференции «Управление развитием крупномасштабных систем» (MLSD'2013). Том II. М.: ИПУ РАН, 2013. С. 312-315.

[4]. Reynolds D.A., Rose R.C. Robust test-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing. 1995. No. 3(1). P. 72-83.

[5]. Guoshen Yu. Solving Inverse Problems with Pricewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity // IEEE Transactions on Image Processing. 2012. No. 21(5). P. 2481-2499.

[6]. Chunxia Guo, Xuehong Qiu. Study of MFCC Speaker Recognition. IT Age. 2005. No. 11. P. 52-56.

[7]. Dempster A., Laird N. Rubin D. Maximum likelihood from incomplete data via the EM algorithm // J. Royal Stat. Soc. 1977. Vol. 39. P. 1–38.

[8]. Baum L. et al. Ann. Math Stat.. 1970, vol. 41, pp. 164–171. DOI: 10.1214/aoms/1177697196.

[9]. Косарев Ю.А., Ли И.В., Ронжин А.Л., Savage J. Методы понимания речи и текста // Труды СПИИРАН. 2002. Т. 2. Вып. 1. С. 157-195.