Распознавание пола диктора на основе GMM-модели голоса

Ромашкин Ю.Н.,
кандидат технических наук
Петров Ю.О.

В статье рассматривается задача автоматического распознавания пола говорящего. С учётом потребности в анализе речевых сигналов относительно малой длительности предложен алгоритм решения, основанный на использовании GMM-модели голоса. Излагаются результаты экспериментальной оценки эффективности алгоритма применительно к речевым сообщениям, полученным в каналах сотовой связи стандарта GSM.

Введение

Методы автоматического распознавания речи находят всё более широкое применение в системах оказания услуг телефонной связи, туристического и гостиничного бизнеса, в технических средствах автоматических информационно-справочных служб и доступа к информации, персонализированной для каждого клиента. Они предназначены для использования произвольным абонентом, не требуют предварительного обучения и являются поэтому дикторонезависимыми. Применительно к условиям приёма речи по проводным линиям телефонной связи общего или внутрикорпоративного пользования существующие методы обеспечивают достаточно хорошую точность распознавания. Однако при использовании абонентом, например, аппаратов мобильной связи, точность распознавания может заметно снижаться, что обусловлено как воздействием помех в радиоканале, так и специфическими искажениями речи при её низкоскоростном кодировании.

Одним из возможных путей повышения эффективности методов автоматического распознавания речи может быть адаптация их параметров по гендерному признаку, т.е. в зависимости от пола говорящего. Заметные различия, например, в частоте основного тона и кратковременном спектре речи мужчин и женщин установлены в [1].

В [2] был предложен способ определения пола диктора по результатам сравнения выборочных плотностей распределения вероятностей, характеризующих значения основного тона. В целом он обеспечил хорошие результаты: вероятности правильного распознавания мужского и женского голосов составили 94,7 и 95,9 % соответственно. Однако такой подход предъявляет повышенные требования как к длительности речи (порядка 1 минуты) для минимизации статистиче-

ской погрешности оценивания выборочной плотности распределения, так и к качеству речи вследствие недостаточной помехоустойчивости оценки основного тона.

В настоящее время экспериментально доказана высокая эффективность применения GMM-метода (модель гауссовской смеси) в различных задачах речевой акустики, включая автоматическое распознавание речи, распознавание языка речевого сообщения и идентификация личности по голосу [3]. Используемые в этом методе мел-кепстральные коэффициенты обладают повышенной помехоустойчивостью и позволяют принимать достоверные решения на относительно коротких интервалах анализа речи. Поэтому интерес представляет экспериментальная оценка эффективности применения данного метода к задаче автоматического распознавания пола диктора по голосу.

1. Математическая формулировка задачи

Задача автоматического распознавания пола диктора по голосу заключается в сопоставлении некоторого речевого сообщения определённому полу диктора. Математически она может быть рассмотрена в рамках теории принятия статистических решений и сформулирована в виде проверки двух альтернативных гипотез.

Пусть задано пространство состояний, включающее две независимые последовательности *N*-мерных векторов $\vec{Y}_M(t)$ и $\vec{Y}_F(t)$ информативных признаков, характеризующие в среднем особенности мужских и женских голосов. А также образовано пространство наблюдений, состоящее из *K* записей $x_i(t) = s_i(t) + \xi_i(t),\ i = \overline{1,K},\ t = \overline{0,T},$ речевых сигналов $s_i(t)$ произвольных дикторов, принятых на фоне помех $\xi_i(t)$. Задача автоматического распознавания пола диктора по голосу состоит в установлении принадлежности каждого наблюдаемого сигнала $x_i(t)$ одному из двух возможных полов.

Переходя от реализации случайного процесса $x_i(t)$ к одноименному *N*-мерному вектору признаков $\vec{X}_i(t)$, получим следующую эквивалентную систему для проверки двух альтернативных статистических гипотез:

$$\begin{cases} H_0 : p[\vec{X}_i(t)] = p[\vec{Y}_M(t)], \\ H_1 : p[\vec{X}_i(t)] = p[\vec{Y}_F(t)]\ ,\ i = \overline{1,K}, \end{cases}$$

т.е. компоненты наблюдаемого вектора признаков принадлежат одному из двух генеральных распределений.

Используем в качестве информативных признаков кратковременные мел-кепстральные коэффициенты и применим аппроксимацию их выборочных распределений с помощью взвешенной суммы M нормальных плотностей распределения с неизвестными параметрами (GMM-модель):

$$p(\vec{X}_i \mid \lambda) = \sum_{j=1}^{M} a_j p(\vec{X}_i \mid \lambda_j),$$

где $p(\vec{X}_i \mid \lambda_j),\ j = \overline{1, M}$, — базисные нормальные плотности распределения этих коэффициентов, a_j — вес *j-й* базисной плотности. Весовые коэффициенты имеют ограничение $\sum_{j=1}^{M} a_j = 1$. Каждая базисная плотность является *N*-мерной гауссовой функцией

$$p(\vec{X}_i \mid \lambda_j) = \frac{1}{(2\pi)^{N/2} \mid D_j \mid^{1/2}} \exp\left\{-\frac{1}{2}(\vec{X}_i - \vec{\mu}_j)^* D_j^{-1}(\vec{X}_i - \vec{\mu}_j)\right\}$$

с вектором $\vec{\mu}_j$ средних значений (размерностью *N*) и ковариационной матрицей D_j (в общем случае размерностью $N \times N$).

Параметры GMM-модели представляются в следующем виде:

$$\lambda_j = \{a_j, \vec{\mu}_j, D_j\},\ j = \overline{1, M}.$$

Они характеризуют индивидуальные особенности голоса каждого диктора и подлежат оцениванию по обучающей реализации речевого сигнала. Нахождение значений параметров GMM-модели голоса диктора, которые наиболее точно отражают выборочные распределения векторов признаков, осуществляется с помощью алгоритма K-средних и EM-алгоритма [4]. Сформируем таким образом средние по множеству обучающих реализаций речевых сообщений GMM-модели мужских (λ_M) и женских (λ_F) голосов.

При статистической независимости последовательности векторов признаков, наблюдаемой на интервале *T*, получим следующее выражение для логарифма функции правдоподобия:

$$P(\vec{X}_0, \vec{X}_1, ..., \vec{X}_T \mid \lambda) = -\ln \prod_{t=0}^{T} p(\vec{X}_i(t) \mid \lambda_j) = -\sum_{t=0}^{T} \ln p(\vec{X}_i(t) \mid \lambda_j).$$

Наконец, применяя критерий максимума апостериорной вероятности, решение о соответствии наблюдаемой последовательности одной из моделей λ_M или λ_F можно записать в следующем виде:

$$R[x_i(t)] = \max[P(\vec{X}_0, \vec{X}_1, ..., \vec{X}_T \mid \lambda_M), P(\vec{X}_0, \vec{X}_1, ..., \vec{X}_T \mid \lambda_F)]. \quad (1)$$

2. Описание обучающей базы речевых сообщений

База данных речевых сообщений, использованных для создания средних GMM-моделей мужских и женских голосов, содержала цифровые записи (при частоте дискретизации 11025 Гц и 16-битном квантовании) телефонных переговоров абонентов сотовой связи стандарта GSM. Возраст абонентов составлял от 21 до 55 лет с примерно равномерным их распределением по трём возрастным группам: 21–30, 31–40 и 41–55 лет.

Для создания средних моделей λ_M и λ_F в экспериментах использовались записи (126 для мужчин и 30 для женщин) речи различных абонентов суммарной длительностью примерно 50 минут с предварительно удалёнными паузами.

В качестве информативных признаков, характеризующих индивидуальные особенности голоса абонента, использовались следующие акустические параметры речи:
— мел-кепстральные коэффициенты (*C*);
— первые производные мел-кепстральных коэффициентов (ΔC);
— вторые производные мел-кепстральных коэффициентов ($\Delta^2 C$).

3. Условия проведения экспериментов

Существующая практика применения GMM-модели в различных задачах обработки речи не даёт чётких рекомендаций о парциальных вкладах используемых параметров *C*, ΔC и $\Delta^2 C$ в общую эффективность алгоритма обработки. В большинстве исследований по умолчанию используются все три указанных информативных признака в предположении их априорной равнозначности.

В проведённых экспериментах последовательно вычислялись два варианта GMM-модели, объединяющие признаки (*C*, ΔC) и (*C*, ΔC, $\Delta^2 C$) соответственно. Парциальные вклады каждого признака оценивались в линейном приближении из условия максимизации вероятности P_D правильного распознавания пола диктора:

$$P_D = \max_{0\le\alpha\le1}[\alpha P_1(C) + (1-\alpha)P_2(\Delta C)],$$
$$P_D = \max_{0\le\beta\le1}\{\beta[\text{a } P_1(C) + (1-\beta)P_2(\Delta C)] + (1-\beta)P_3(\Delta^2 C)\}, \quad (2)$$

где P_1, P_2 и P_3 — оценки вероятности правильного распознавания, получаемые при раздельном использовании каждого из признаков.

Вычисление мел-кепстральных коэффициентов и их производных проводилось для сегментов речевых сигналов постоянной длительности 12 мс. с использованием стандартных функций среды Matlab 7. Размерность *N* соответствующих векторов равнялась 16. Матрица D_j имела диагональный вид (т.е. компоненты вектора признака считались статистически независимыми).

4. Результаты экспериментов

В экспериментах по оценке эффективности алгоритма автоматического распознавания использовались 90 тестовых записей речевых сообщений абонентов-мужчин и 30 записей женщин, полученных в канале сотовой связи стандарта GSM. При этом записи, использованные на этапах обу-

чения и тестирования алгоритма, не перекрывались как по составу абонентов, так и по времени. Длительности речевых сообщений при тестировании составляли 10 и 5 секунд (с автоматически удалёнными паузами).

Сначала выбирался порядок GMM-модели, равный 4, и для него находилось оптимальное значение весового коэффициента α, удовлетворяющее первому уравнению в (2). Результаты этих экспериментов при длительности тестовых сигналов T=10 и 5 секунд представлены графически на рис. 1 и 2 соответственно в виде зависимостей вероятности правильного распознавания пола диктора от α.

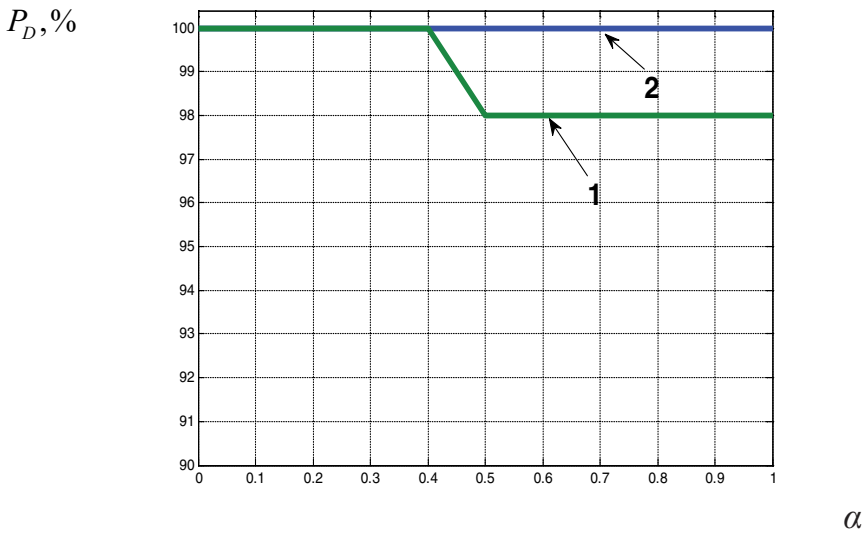

Рис. 1. Экспериментальные оценки точности распознавания пола диктора в зависимости от коэффициента α (T=10 с.): 1 — для женских голосов; 2 — для мужских голосов

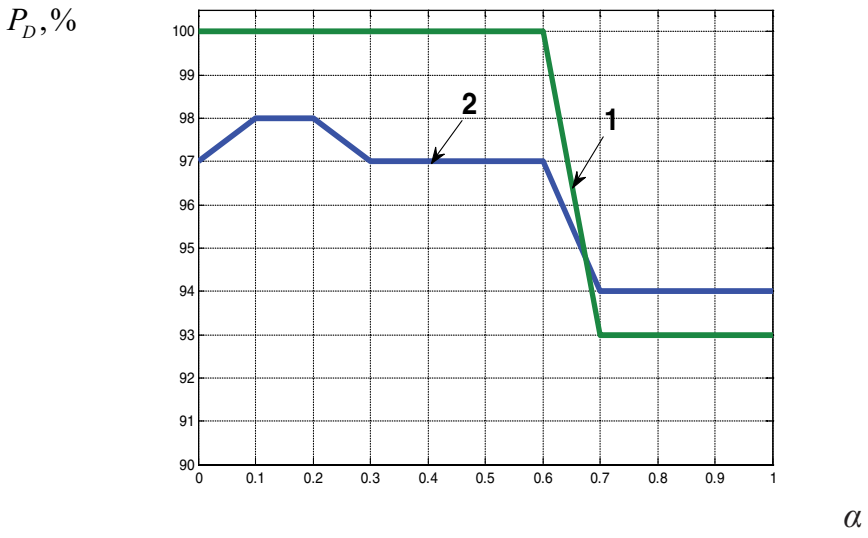

Рис. 2. Экспериментальные оценки точности распознавания пола диктора в зависимости от коэффициента α (T=5 с.): 1 — для женских голосов; 2 — для мужских голосов

Из полученных результатов следует, что при T=5–10 секунд наиболее рациональными являются значения α=0,1–0,2, т.е. относительный вклад мел-кепстральных коэффициентов в суммарную эффективность алгоритма оказался существенно меньшим, чем вклад их первых производных. Данный эффект можно объяснить тем, что в реализованном алгоритме к вектору мел-кепстральных коэффициентов не применялись известные методы нормализации [3], поэтому функции компенсации амплитудно-частотных характеристик каналов приёма и передачи речи, изменяющегося расстояния до микрофона мобильного телефона и аддитивных помех в радиоканале в этом случае переносятся на первые производные коэффициентов.

При указанных выше значениях α женские голоса распознаются алгоритмом безошибочно при T=10 и 5 с., а мужские — также безошибочно при T=10 с. и с ошибкой, равной 2 %, при T=5 с. При аппроксимации результатов испытаний биномиальным распределением доверительные интервалы полученных оценок вероятности правильного распознавания при T=5 с. и коэффициенте доверия 0,95 составили (91,5–100) % для женских голосов и (93,1–99,6) % для мужских [5]. Более узкий доверительный интервал для последних является следствием того, что тестовая выборка записей для мужских голосов оказалась в экспериментах более представительной (126 записей), чем для женских (30 записей).

Далее оптимизировалось значение весового коэффициента β в соответствии со вторым уравнением в (2) при фиксированном α=0,2. Результаты этих экспериментов представлены графически на рис. 3 и 4 в виде аналогичных зависимостей $P_D(\beta)$. Они показывают, что наиболее рациональными можно считать значения β=0,6–1,0. Однако добавление второй производной мел-кепстральных коэффициентов по времени практически не повышает эффективность распознавания, требуя при этом дополнительного времени обработки. Алгоритм по-прежнему безошибочно распознал все женские голоса при T=10 и 5 с., а также мужские

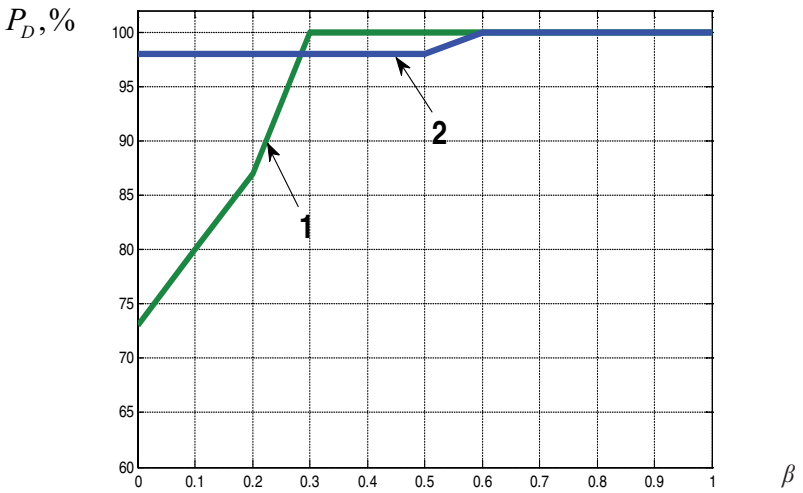

Рис. 3. Экспериментальные оценки точности распознавания пола диктора в зависимости от коэффициента β (T=10 с.): 1 — для женских голосов; 2 — для мужских голосов

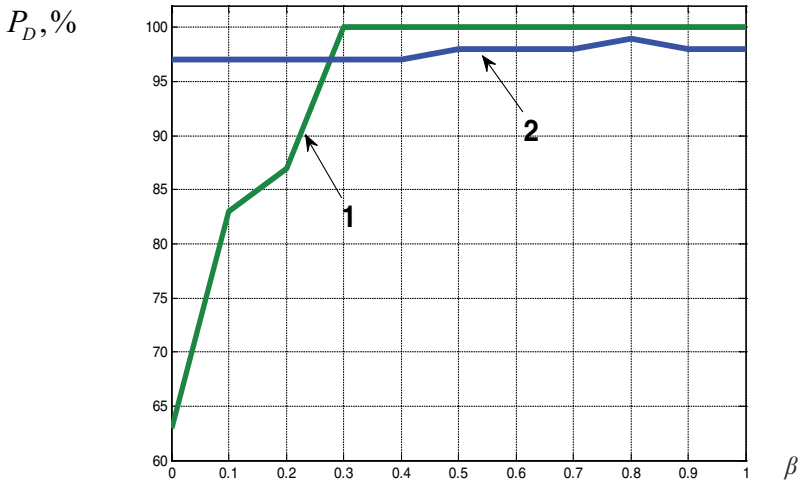

Рис. 4. Экспериментальные оценки точности распознавания пола диктора в зависимости от коэффициента β (T=5 с.): 1 — для женских голосов; 2 — для мужских голосов

при T=10 с. Ошибка распознавания мужских голосов при T=5 с. не уменьшилась и составила 2 %.

Очевидный интерес представляет поиск возможностей повышения точности распознавания при малых (T=5 с.) длительностях речевых сообщений за счёт увеличения порядка используемой GMM-модели. Результаты проведённых экспериментов (с использованием мел-кепстральных коэффициентов и их первых производных при α=0,2) показывают, что уже при увеличении порядка GMM-модели в 2 раза (до M=8) алгоритм обеспечил безошибочное распознавание пола абонентов для всех тестовых записей как с мужскими, так и женскими голосами. Доверительный интервал полученных оценок вероятности правильного распознавания мужских голосов в этом случае составил (96,7–100) % при коэффициенте доверия 0,95.

Заключение

Алгоритмы на основе GMM-модели речи могут успешно применяться для решения различных задач обработки речевой информации, в том числе автоматического распознавания пола говорящего. Такой подход наряду с повышенной эффективностью распознавания позволяет снизить требования к длительности анализируемого речевого сигнала. Полученные экспериментальные результаты показывают возможность надёжного распознавания пола при анализе коротких речевых сообщений абонентов сотовой телефонной связи стандарта GSM.

Литература

1. Михайлов В.Г., Златоустова Л.В. Измерение параметров речи. М.: Радио и связь, 1987. 168 с.

2. Сорокин В.Н., Макаров И.С. Определение пола диктора по голосу // Акуст. журнал, 2008. Т. 54. № 4. С. 659–668.

3. Benesty J., Sondhi M., Huang Y. Springer Handbook of Speech Processing, 2008. 1176 p.

4. Аграновский А.В., Леднов Д.А. Теоретические аспекты алгоритмов обработки и классификации речевых сигналов. М.: Радио и связь, 2004. 164 с.

5. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука, 1983. 416 с.

Ромашкин Юрий Николаевич
кандидат технических наук. Московский государственный институт радиотехники, электроники и автоматики (технический университет)

Петров Юрий Олегович
Государственное учреждение «Войсковая часть 35533».