В. Я. Чучупал, На пути к созданию робастных методов распознавания речи, *ИТиВС*, 2004, выпуск 2, 46–51

Инвариантность методов распознавания речи*

В.Я. Чучупал

Аннотация. Рассматриваются различные варианты адаптации параметров системы распознавания речи к голосу диктора и характеристикам окружающей среды. Приведены результаты исследований метода адаптации на основе использования кластеров акустических моделей. Предложен способ формирования кластеров на основе алгоритма кластеризации данных "снизу вверх" и использования меры близости на основе оценок максимального правдоподобия. Показано, что использование кластерных моделей в системе распознавания речи позволяет увеличить точность распознавания и может рассматриваться как альтернативный либо дополнительный, по отношению с существующим, способ подстройки параметров системы в процессе ее функционирования.

Проблема адаптации к голосу и окружающей среде

Современные системы распознавания речи основаны на статистических моделях, которые описывают акустические свойства звуков речи. Параметры акустических моделей оцениваются на материале речевых корпусов данных, которые, в свою очередь, специальным образом проектируются, записываются и аннотируются.

В реальных практических ситуациях речевой сигнал наблюдается в условиях существенно отличных от тех, в которых проводился сбор корпуса данных. Это несоответствие между обучающим и реальным речевым материалом обусловлено целым рядом факторов, в частности:

- наличием и характером окружающего шума;
- различием в используемых микрофонах или каналах связи;
- изменением характеристик голоса диктора (изменения личности говорящего, его эмоционального и физического состояния).

На рисунке схематически представлена достаточно распространенная среда функционирования системы распознавания речи.

Шум, даже относительно небольших уровней, может существенно, если не критически, влиять на точность работы системы распознавания речи.

Помимо непосредственного искажающего воздействия на акустический сигнал, шум, при достаточно высокой его интенсивности, влияет и на источник речевого сигнала. Так называемый *ломбард-эффект* заключается в том, что в условиях сильных помех говорящий изменяет громкость речи с тем, чтобы обеспечить достаточно надежное её распознавание. Изменение параметров в таком случае включает довольно сложные реорганизации речеобразующей системы, которые в конечном итоге существенно изменяют характеристики голоса. Таким образом, речь воспроизводимая в шумной среде, существенно отличается по характеристикам от речи, воспроизводимой тем же лицом в спокойной обстановке.

Для многих практически очень важных приложений систем распознавания речи наличие перцептивно заметного внешнего шума (движение автомобиля, наличие мешающих дикторов), изме-

* Работа выполнялась при поддержке грантов РФФИ 02-01-00453 и 04-01-00588

нений характеристик канала передачи (телефонный канал, тип микрофона и его расположение по отношению к говорящему) или голоса диктора является достаточно типичным.

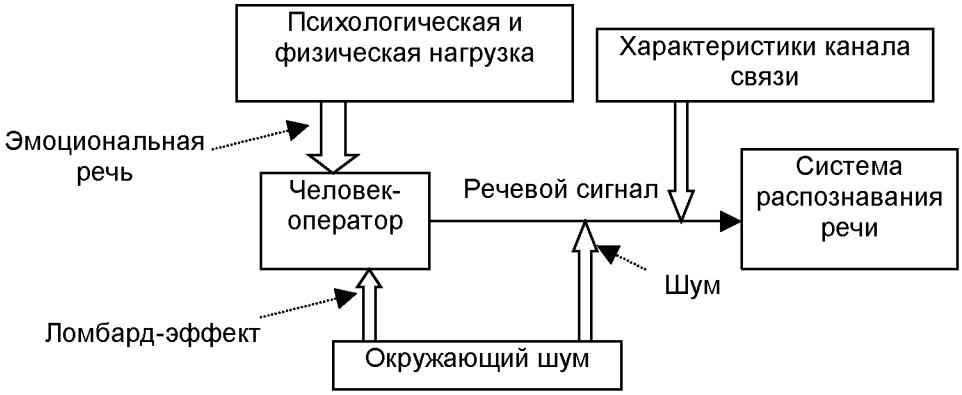

Рис. 1. Источники вариативности в системе распознавания речи

В связи с этим проблема обеспечения надежной работы систем распознавания речи в реальных акустических условиях сейчас одна из наиболее важных в этой области.

Принято считать систему распознавания речи робастной, если эта система сохраняет работоспособность для широкого круга приложений. В частности, робастность подразумевает возможность успешного использования в условиях шума, стресса, искажений в канале связи, а в более широком смысле и работу с естественной разговорной речью

Современные методы построения робастных систем распознавания речи

К настоящему времени разработано большое число способов повышения надежности систем распознавания речи.

Если не рассматривать специальных технических средств сбора речевых данных (многоканальные и мультимодальные интерфейсы, ларингофоны и т.п.), адаптация в существующих системах распознавания речи осуществляется одним из следующих способов: использованием более робастных систем параметров и мер сходства, предобработкой речевого сигнала на входе системы распознавания речи либо модификацией (адаптацией) параметров акустических моделей звуков.

Например, общепринятым способом повышения точности распознавания в присутствии помех является дополнение системы кратковременных параметров их первыми и вторыми производными по времени.

Для речевых команд в условиях ломбард-эффекта речи и сильных акустических помех применение производных параметров приводит в зависимости от уровня шума и характера искажений к увеличению точности распознавания на 15% - 50%.

Предварительная обработка речевого сигнала включает в себя методы его нормализации, такие как вычитание кепстрального или спектрального среднего, фильтрация временных траекторий параметров (RASTA – обработка), методы нейронных сетей, а также калмановская фильтрация и различные модификации спектрального вычитания для снижения влияния квазистационарных аддитивных помех.

Спектральное вычитание [1,2] и методы на его основе применяются наиболее часто, так как легко реализуется в виде дополнения к уже существующей системе распознавания.

Если наблюдаемый зашумленный сигнал $x(t)$ может быть представлен как сумма "полезного" речевого сигнала $s(t)$ и шума $n(t)$:

$$x(t) = s(t) + n(t),$$

то для метода спектрального вычитания плотность мощности $P_s(iw)$ полезного сигнала $s(t)$ оценивается как:

$$P_s(iw) = P_x(iw) - P_n(iw).$$

В условиях кратковременной поблочной обработки данных кратковременный амплитудный спектр сигнала $|S(t,iw)|$ (на основе которого, как правило, вычисляются параметры) оценивается как

$$|S(t,iw)|^2 = \begin{cases} |X_i(t,iw)|^2 - A(t)|N(t,iw)|^2, \text{если } |\mathrm{X_i(t,iw)}|^2 \geq (\mathrm{A(t)+B})|\mathrm{N(t,iw)}|^2 \\ B|N(t,iw)|^2, \end{cases}$$

где $|X(t,iw)|$ и $|N(t,iw)|$ - амплитудные спектры наблюдаемого сигнала и шума, соответственно.

Коэффициент A(t) - фактор переоценивания, вообще говоря, зависит от соотношения сигнал/шум на сегменте анализа и имеет типичные значения, близкие к 0.7 - 0.95, а коэффициент B - спектральный порог - выбирается в диапазоне 0.01 - 0.1.

Стандартный выигрыш в точности при использовании подобного подхода составляет для распознавания команд в условиях движущегося автомобиля - от 8 до 13%.

Для подстройки на голос диктора и характеристики канала связи широко используются также модификация или адаптация параметров акустических моделей звуков.

Для марковских моделей с непрерывными плотностями вероятности, вероятность наблюдения (вектора параметров) x для каждого состояния модели представлена суммой некоторого (например, M) числа взвешенных нормальных распределений:

$$p(x) = \sum_{k=1}^{M} c_{jk} N(x, \mu_{jk}, \Sigma_{jk}),$$

где c_{jk} – веса распределений (c_{jk}>=0, Σc_{jk}=1), а μ и Σ - вектор средних и ковариационная матрица соответственно.

Модификация моделей в таком случае сводится для метода линейной регрессии и максимального правдоподобия (maximum likelihood linear regression), сокращенно MLLR [3], к подстройке вектора средних для нормальных распределений и/или модификации ковариационной матрицы:

$\boldsymbol{\mu} = \boldsymbol{A\mu} + \boldsymbol{b}$,

$\boldsymbol{\Sigma}' = \boldsymbol{H}\,\boldsymbol{\Sigma}\,\boldsymbol{H}^t$,

где $\boldsymbol{H}$ и $\boldsymbol{A}$ – трансформационные матрицы.

При использовании стохастического аддитивного преобразования (stochastic additive transform – SAT) ковариационная матрица плотности - диагональная, а матрица аффинного преобразования средних – единичная, то есть в скалярной адаптация может быть записана как:

$\mu'_i = \mu_i + \mu_{bi}$

$\delta'^2_{ii} = \delta^2_{ii} + \delta^2_{bii}$.

Применение MLLR для адаптации к голосу диктора в задаче распознавании слитной речи на большом словаре привело к снижению абсолютной ошибки распознавания (для материала корпуса данных Wall Street Journal) с 7.7% до 6.6% (относительная ошибка тем самым снизилась на 13%).

Адаптация к голосу и среде на основе кластерных моделей

В течение последних лет в ВЦ РАН исследуется метод адаптации системы распознавания речи к голосу дикторов и влиянию окружающей среды, который основан на использовании кластеров акустических моделей. В частности, рассматривалась процедура кластеризации голосов дикторов, синтеза кластерных акустических моделей и использования таких моделей в системе распознавания речи.

Предпосылками к этому подходу были следующие наблюдения:

- успешное применение эмпирической техники раздельного акустического моделирования мужских и женских голосов в системах распознавания речи;
- оценка параметров акустико-фонетических моделей непосредственно в естественной, шумной среде обычно приводит к существенно лучшим результатам по точности распознавания речи, нежели рассмотренные выше методы адаптации;
- вариативность речевого сигнала довольно сложно описать аналитически.

Кластерные акустико-фонетические модели названы так потому, что в таком случае звук имеет несколько акустико-фонетических моделей, которые являются специализированными, например, по отношению к среде и дикторам и используются при распознавании речи одновременно и конкурентным образом.

Задачами проводимых исследований, в частности, являются:

- построение систематического метода кластеризации статистических моделей звуков по акустическим характеристикам голоса и окружающей среды;
- построение методов экономичной интеграции кластерных моделей в процедуры распознавания речи (в частности, исследование возможности обойтись без использования детекторов кластера);
- выяснение возможности автоматического определения кластера модели по входному сигналу.

Построение кластерных акустических моделей

Кластеризация акустических моделей выполняется методом построения кластеров снизу – вверх. Алгоритм вычислений имеет следующий вид.

Пусть задано множество диктров $s_i, \ldots s_N$.

1. Сформируем начальное множество кластеров голосов дикторов $S_i \ldots S_N$ как кластеров, состоящих из голоса одного диктора, $S_i = \{s_i\}$.

2. Для каждого кластера S_i соберем соотвествующую выборку речевых параметров и оценим модель (плотность распределения) параметров M_i.

Определим "качество" данных кластера S_i на модели M_j как величину правдоподобия данных:

$L(S_i \mid M_j) = \sum_{o_j \in S_i} \log P(o_j \mid M_j)$, вычисленного на всех наблюдениях o_j из кластера S_i.

Пусть S_{i+j} – кластер, образованный слиянием кластеров с индексами i и j. Определим "выигрыш" (или расстояние между кластерами *Si u Sj*) от операции слияния кластеров *Si* и *Sj* как изменение в правдоподобии данных в результате замены двух кластеров их обьединением:

$L(S_i, S_j) = L(S_{i+j} \mid M_{i+j}) - (L(S_i|M_i) + L(S_j|M_j))$.

3. Вычислим попарные расстояния $L(S_i, S_j)$ между всеми кластерами и выберем пару кластеров *Si* и *Sj*, для которых такое расстояние минимально. Если величина расстояния $L(S_i, S_j)$ оказалась больше некоторого заранее выбранного порога или число кластеров N меньше минимального - переходим на шаг 5.

В противном случае обьединим эти кластеры (родители) *Si* и *Sj* в один общий S_{i+j} и заменим этим новым кластером родительские. Текущее число кластеров N уменьшится на 1.

4. Оценим для нового кластера S_{i+j} его модель M_{i+j}. Перейдем на шаг 3.

5. Закончим вычисления и сохраним результаты кластеризации

Описанная процедура формирования кластеров может приводить к "переобучению" моделей, поэтому исходные данные для всех дикторов перед началом работы алгоритма разбивались на два равномощных множества – одно использовалось в соответствии с алгоритмом для формирования кластеров, другое служило для подтверждения правильности выбора объединяемой пары кластеров.

Поскольку число векторов-параметров для каждого диктора было примерно (с точностью до 50%) одинаковым, в нормировке по размеру выборки необходимости не возникало.

Результаты численных экспериментов

Численные эксперименты заключались в построении кластеров голосов дикторов на материале речевого корпуса данных, оценивании параметров кластерных моделей, измерении точности распознавания в зависимости от числа кластеров и правильности определения кластера, к которому относится запись голоса диктора.

Для обучения (оценки параметров распределений) использовался материал фонетической части речевой базы данных TeCoRus [4]. Для "доводки" моделей привлекался числовой материал того же речевого корпуса данных.

Предобработка речевого сигнала заключалась в вычислении набора из 15 мел-кепстральных коэффициентов и их первых производных по времени (дельта-кепстра). Система распознавания была сконфигурирована для представления речевого сигнала помощью марковских моделей с дискретными плотностями вероятности. В качестве кодовых книг (формировались раздельные книги для кепстра и дельта-кепстра) использовались самоорганизующиеся карты признаков.

Оценка правдоподобия вычислялась следующим образом. Если x – вектор параметров сигнала в некоторый момент времени и y – соответствующий ему элемент кодовой книги, то вероятность p(x|M) оценивалась как частота: $p(x \mid M) = \frac{N_y}{\sum_{v=1}^{529} N_v}$, где N_v – число, показывающее сколько раз, когда элемент книги с индексом v встречался за время обучения.

В данном случае использовалась взвешенная оценка, взятая по частотам, соответствующим четырем ближайшим элементам кодовой книги.

Фактически оценивалось свыше 1600 распределений, соответствующих примерно 540 "физических" моделей аллофонов. Соответствие между логическими и физическими моделями устанавливалось с помощью бинарного решающего дерева. При синтезе лексической сети кроссворды (модели межсловных звуков) не использовались. Не применялась и модель языка.

Предварительно эмпирически было отмечено, что формировать кластеры участием менее 15 дикторов оказалось явно нецелесообразно, так как акустические модели оценивались грубо и точность распознавания, в результате, ухудшалась.

Для построения кластеров дикторов использовался материал фонетически сбалансированных фраз из TeCoRus от 50 дикторов, 25 мужчин и 25 женщин. При этом никакой предварительной фонетической сегментации или классификации векторов параметров не делалось – весь набор векторов параметров от одного диктора, включая участки пауз, использовался для построения одной "общей" модели голоса диктора.

Для измерений использовались цифры и последовательности цифр от 10 дикторов, не входящих в обучающее множество. Дикторы выбирались случайным образом, однако сохранялось равное соотношение между мужчинами и женщинами. От каждого диктора использовалось 10 последовательностей (2-,3- и 6-значных) цифр и 10 раздельно произносимых цифр.

Первоначальное разделение дикторов на два кластера привело к повышению точности распознавания цепочек цифр почти на 5%, однако использование трех кластеров не улучшило точность распознавания. Для изолированно произносимых чисел эффект кластеризации был незаметен, так как точность распознавания уже на одном кластере была выше 99%.

Уменьшение точности для случая трех кластеров связано, по-видимому, только с уменьшением обучающей выборки – все ошибки распознавания относились к "мужскому"кластеру, который был разделен на два, соответственно, с уменьшением размера обучающих данных.

Исследовалась также точность определения кластера диктора в процессе распознавания. Оказалось, что она существенно превышает первоначальные оценки, то есть о типе голоса (для двух - и трех - кластерных моделей) можно довольно точно судить по сравнительно коротким сегментам речевого материала.

Точность определения кластера диктора также оценивалась на материале из изолированно произносимых цифр и слитно произносимых последовательностей из цифр. При распознавании кластера каждое отдельное произнесение цифры или последовательности цифр рассматривалось как отдельная попытка.

В этом случае для двух - кластерных моделей точность распознавания кластера голоса диктора превышала 99%, а для трех - кластерных моделей – 98%. Более того, даже для раздельно произносимых цифр (длительность высказывания менее 0.5 секунды) двух - кластерные модели распознавались примерно с такой же точностью, а трех - кластерные модели – существенно хуже – около 85%.

Один из выводов, сделанных в результате анализа полученных кластеров, состоял в следующем. При разбиении голоса дикторов на два класса решающую роль (для дикторов от 16 до 65 лет) играл пол диктора – полученные кластеры полностью (на 100%) совпадали с интуитивно понятными кластерами, образованными из голосов мужчин и женщин. Дальнейшее разбиение дикторов проходило внутри кластеров, причем здесь уже существенную роль играли не только характеристики голоса диктора, но и внешние условия записи. Так, при переходе к 3- и 4 –кластерным моделям происходило разбиение мужских и женских голосов на два кластера, соответствующие типам использованных микрофонов.

Заключение

Рассмотрен вариант адаптации системы распознавания речи к голосу диктора, который основан на использовании кластеров акустических моделей, оцениваемых на основе стандартных процедур максимума правдоподобия.

Приведены алгоритм группирования акустических моделей и результаты численных экспериментов по построению кластеров на материале речевого корпуса данных, оцениванию параметров кластерных моделей, измерению точности распознавания в зависимости от числа кластеров и правильности определения кластера, к которому относится запись голоса диктора.

Показано, что реализация кластерных моделей позволяет увеличить точность распознавания. В случае использования двух кластеров полученные множества голосов дикторов совпадают с традиционным разделением на мужские и женские голоса. В процессе распознавания кластер диктора также может быть идентифицирован с высокой точностью по небольшому участку речевого сигнала.

Литература

1. Boll S.F. Suppression of Acoustic Noise in Seech Using Spectral Subtraction//IEEE Trans. ASSP, Vol.27, No.2, pp 113 - 120, 1979
2. Sondhi M.M., Schmidt C.E., Rabiner L.R. Improving the Quality of Noisy Speech Signal // Bell Syst. Tech Journ, Vol.60, No.8. 1981, pp.1847-1858.
3. Leggetter, C.J. Woodland P.C. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs // Computer Speech and Language Journal, 9, 171-186, 1995
4. Kouznetsov V., Chuchupal V, Makovkin K., Chichagov A. Design and implementation of Russian telephone speech database // Proc. Int Workshop “Speech and Computer”, SPECOM-99/ Moscow, 1999, pp.179-181.

Чучупал Владимир Яковлевич. Кандидат физико-математических наук. Автор нескольких десятков научных и методических публикаций по проблемам обработки и распознавания речи и биометрической идентификации личности. Область научных интересов: математическое моделирование и обработка речевых сигналов, распознавание естественной речи, идентификация личности по голосу, выделение ключевых слов в потоке слитной речи. Руководитель сектора цифровой обработки и распознавания речи ВЦ РАН.