ВЫДЕЛЕНИЕ ОСНОВНОГО ТОНА, ОСНОВАННОЕ НА МАРКОВСКОЙ МОДЕЛИ

ЛЮДОВИК ЕВГЕНИЙ КУЗЬМИЧ

Институт кибернетики им. В.М.Глушкова АН УССР
Киев, СССР 252207

АННОТАЦИЯ

Учет основных свойств траекторий основного тона на базе марковской модели позволяет существенно повысить надежность определения периода основного тона и признака тон/шум.

ВВЕДЕНИЕ

Чтобы повысить надежность выделения основного тона /ОТ/ предлагается применить к этой задаче статистическую теорию распознавания образов. Рассматриваемая в связи с этим статистическая модель включает две компоненты:

1) модель, описывающую зависимость принимаемого первичного описания речевого сигнала от значения признака тон/шум и периода ОТ на отдельном интервале анализа, и
2) модель, отражающую основные свойства траекторий признака тон/шум и периода ОТ на фрагментах речевого сигнала из многих интервалов анализа.

Для построения модели разобьем весь диапазон значений ОТ от 62,5 Гц до 500 Гц на 12 сегментов в равномерной логарифмической шкале. При этом на каждую октаву приходится по четыре сегмента. Во временной области получим 12 сегментов $[\tau_{s-1},\tau_s]$, $s=1,\ldots,12$, длина которых достаточно хорошо соответствует максимально возможной разности периодов ОТ на соседних интервалах анализа, сдвиг между которыми равен 10 мс.

Будем исходить из того, что речевой сигнал подвергнут процедуре выравнивания спектра (например, путем обратной фильтрации системой линейного предсказания) и пропущен через фильтр низких частот, после чего вычислена нормированная автокорреляционная функция $R(i)$, $0 \le i \le L$.

Для каждого из 12 сегментов найдем опорное значение ОТ T_s:

$$T_s = \arg\max_{i:\tau_{s-1} \le i < \tau_s} R(i)$$

и в дальнейшем вместо автокорреляционной функции $R(i)$ будем рассматривать ее "проекцию" $\bar{R}(s) = R(T_s), 1 \le s \le 12$. Кроме того, впредь, говоря о периоде ОТ, как правило, будем иметь в виду номер сегмента s, в который попадает этот период. Номер сегмента s является огрубленной характеристикой ОТ; уточненное значение, соответствующее огрубленному значению s, равно опорному значению T_s.

МОДЕЛЬ ОСНОВНОГО ТОНА НА ОТДЕЛЬНОМ ИНТЕРВАЛЕ АНАЛИЗА

Построение такой модели означает определение зависимости плотности совместного распределения вероятностей отсчетов проекции автокорреляционной функции (АКФ) $\bar{R} = \{\bar{R}(s)\}_{s=1}^{12}$ от значений признака тон/шум v и периода ОТ s : $p(\bar{R}/s,v)$, где $v=1$ соответствует признаку "тон", а $v=0$ – признаку "шум".

Задание такого многомерного распределения, зависящего к тому же от параметров, является в общем случае довольно сложной задачей. Здесь, однако, допустимо прибегнуть, например, к такому упрощению. Найдем максимум ПАФ $\bar{R}$:

$$s^* = \arg\max_{s:1 \le s \le 12} \bar{R}(s) .$$

Объявление s^* огрубленным значением периода ОТ соответствует корреляционному методу выделения ОТ; мы, однако, сделаем некоторый шаг вперед по сравнению с этим методом, если учтем вероятности получения различных значений s^* при фактическом значении s : $P_1(s^*/s)$. Причем здесь можно ограничиться заданием вероятности правильной оценки $P_1(s/s)$ и вероятности ошибки, соответствующей удвоенному периоду ОТ, $P_1(s+4/s)$, полагая остальные ошибки равновероятными с вероятностью

$$P_1(s'/s)=(1-P_1(s/s)-P_1(s+4/s))/10,$$
$$s' \ne s,\ s' \ne s+4.$$

Величина $R^*=\bar{R}(s^*)$ в корреляционном методе служит для определения признака тон/шум; при этом решение принимается в результате сравнения этой величины с некоторым порогом. Мы, однако, определим вероятности значений R^* в зависимости от значения признака тон/шум: $p_2(R^*/v)$, сохранив больше информации, чем используется в корреляционном методе.

Теперь статистические модели вокализованного и невокализованного интервалов могут быть определены, например, следующим образом:

$$p(\bar{R}/s,v=1) = p(R^*,s^*/s,v=1) = $$
$$= P_1(s^*/s)\cdot p_2(R^*/v=1) ,$$
$$p(\bar{R}/s,v=0) = p(R^*,s^*/s,v=0) = $$
$$= C\cdot p_2(R^*/v=0),$$

где $C = 1/12$ – вероятность любого конкретного значения s^* на невокализованном интервале.

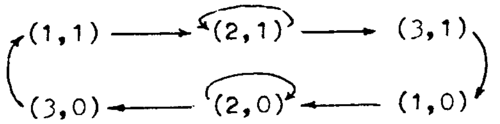
Более точная статистическая модель получится, если учесть зависимость между $\bar{R}(s)$ и $\bar{R}(s+4)$ при фактическом значении ОТ s . Возможны и другие варианты.

МАРКОВСКАЯ МОДЕЛЬ ТРАЕКТОРИЙ ОТ

Модель траекторий ОТ опирается на следующие их свойства. Во-первых, продолжительность вокализованных и невокализованных отрезков речевых сигналов ограничена снизу. Во-вторых, имеет место сильная зависимость между значениями периода ОТ на соседних вокализованных интервалах анализа. Аналогичная зависимость, хотя и в меньшей степени, наблюдается между значениями периода ОТ на интервалах, разделенных невокализованным участком сигнала, причем эта зависимость убывает с увеличением длительности разделяющего невокализованного участка.

Первое из этих свойств хорошо отражается следующей марковской моделью, генерирующей траектории признака тон/шум. Допустим, что минимальная длина вокализованных и невокализованных участков равна 3. Введем по три состояния $(1,v)$, $(2,v)$ и $(3,v)$ для каждого значения признака тон/шум $v=0,1$.

Граф переходов модели имеет вид:



Состояния $(1,1)$ и $(1,0)$ являются начальными при порождении вокализованного и невокализованного участков соответственно, а $(3,1)$ и $(3,0)$ – конечными. Повторение состояний $(2,1)$ и $(2,0)$ позволяет генерировать участки вокализованной и невокализованной речи любой длины, большей или равной трем интервалам анализа. Для полного описания этой модели необходимо задать $P_v(2/2)$ – вероятность перехода по петле в каждое из этих повторяемых состояний, при этом $P_v(3/2) = 1 - P_v(2/2)$.

Второе из сформулированных выше свойств также может быть описано с помощью некоторой марковской модели; эта модель будет генерировать траектории значений периода ОТ. Модель содержит I2 состояний, соответствующих огрубленным значениям ОТ s. С достаточно хорошим приближением можно считать, что огрубленные значения периода ОТ s и s' на соседних интервалах анализа могут отличаться не более, чем на единицу: $|s - s'| \leqslant 1$. Задав вероятности переходов $P_3(s'/s)$, определяющие среднюю скорость изменения периода в огрубленной шкале, получаем марковскую модель траекторий периода ОТ.

Объединение модели траекторий признака тон/шум и модели траекторий периода ОТ приводит к совокупной марковской модели траекторий ОТ. Возможные переходы и их вероятности представлены ниже:

$(s,1,1) \rightarrow (s',2,1),\ P_3(s'/s),$

$(s,2,1) \rightarrow (s',2,1),\ P_3(s'/s) \cdot P_v(2/2),$

$(s,2,1) \rightarrow (s',3,1),\ P_3(s'/s) \cdot P_v(3/2),$

$(s,3,1) \rightarrow (s',1,0),\ P_3(s'/s),$

$(s,1,0) \rightarrow (s',2,0),\ P_3(s'/s),$

$(s,2,0) \rightarrow (s',2,0),\ P_3(s'/s) \cdot P_v(2/2),$

$(s,2,0) \rightarrow (s',3,0),\ P_3(s'/s)\ P_v(3/2),$

$(s,3,0) \rightarrow (s',1,1),\ P_3(s'/s),$

причем во всех случаях $|s - s'| \leqslant 1$.

Состояния с различными значениями s для невокализованных интервалов вводятся, чтобы отразить зависимость значений периода ОТ на вокализованных интервалах, примыкающих с двух сторон к невокализованному участку.

Введем следующие обозначения:

$S_n = (s_n, i_n, v_n)$ – состояние модели на n-м интервале,

R_n – ПАФ сигнала на n-м интервале;

$p(R_n/S_n)$ – условная плотность вероятности параметров ПАФ в зависимости от значения признака тон/шум и периода ОТ;

$P(S_{n+1}/S_n)$ – вероятность перехода из состояния S_n в состояние S_{n+1};

$P(S_1/S_0)$ – распределение вероятностей начального состояния.

В качестве $P(S_1/S_0) = P(S_1)$ естественно взять распределение с равными вероятностями для всех невокализованных состояний и нулевыми вероятностями для вокализованных.

Теперь задача выделения ОТ сводится к задаче определения состояний построенной марковской модели, наилучшим образом соответствующих наблюденному речевому сигналу. Возможны различные постановки этой задачи.

Можно, например, поставить задачу отыскания наиболее правдоподобной траектории ОТ по всей реализации речевого сигнала. Для этого следует максимизировать апостериорную вероятность траектории:

$$P(S_1,\dots,S_N/R_1,\dots,R_N) = \prod_{n=1}^{N} P(S_n/S_{n-1}) \cdot p(R_n/S_n)/p(R_1,\dots,R_N). \qquad \text{(I)}$$

Эта задача решается следующим алгоритмом динамического программирования:

$$\left.\begin{aligned} I(S',n) &= \underset{S \in Q(S')}{\operatorname{argmax}} \left(F(S,n-1) + g(R_n,S,S')\right) \\ F(S',n) &= \max_{S \in Q(S')} \left(F(S,n-1) + g(R_n,S,S')\right) \end{aligned}\right\} 2 \leqslant n \leqslant N,$$

$$S^*_N = \underset{S}{\operatorname{argmax}}\ F(S,N), \qquad \text{(2)}$$

$$S^*_{n-1} = I(S^*_n, n),\ n = N,\dots,2.$$

В этих формулах:

$g(R_n,S,S') = \ln P(S'/S) + \ln p(R_n/S'),$

$F(S,1) = \ln P(S) + \ln p(R_1/S),$

$Q(S')$ – множество состояний, из которых возможен переход в состояние S'.

Другим важным вариантом постановки задачи является минимизация вероятности ошибки выделения ОТ на m-м интервале анализа по фрагменту реализации из первых $m+k$ интервалов. Для решения этой задачи, как известно, следует определить, какое из интересующих нас событий имеет максимальную апостериорную вероятность. Мы будем рассматривать I3 событий: одно, означающее невокализованность m-го интервала, и еще I2, соответствующих двенадцати возможным значениям s_m периода на m-м интервале, $1 \leqslant s_m \leqslant 12$.

Для вычисления вероятностей этих событий следует просуммировать апостериорные вероятности состояний модели на m-м интервале, составляющих эти события. Апостериорная вероятность состояния $S_m = (s_m, i_m, v_m)$ определяется следующей формулой:

$$P(s_m,i_m,v_m/R_1,\dots,R_{m+k}) = P(S_m/R_1,\dots,R_{m+k}) = \sum_{\substack{S_1,\dots,S_{m-1},\\ S_{m+1},\dots,S_{m+k}}} (1/A) \cdot \prod_{n=1}^{m+k} P(S_n/S_{n-1}) \cdot p(R_n/S_n),$$

где $A = p(R_1,\dots,R_{m+k})$. Суммирование производится по всем состояниям на первых $m+k$ интервалах, кроме m-го.

Теперь можно вычислить апостериорные вероятности огрубленных значений периода:

$$P(s_m, v_m = 1/R_1,\dots,R_{m+k}) = \sum_{i_m} P(s_m,i_m,v_m = 1/R_1,\dots,R_{m+k}),\quad 1 \leqslant s_m \leqslant 12,$$

и апостериорную вероятность того, что m-й интервал невокализованный:

$$P(v_m = 0/R_1,\dots,R_{m+k}) = \sum_{i_m, s_m} P(s_m,i_m,v_m = 0\ /R_1,\dots,R_{m+k})$$

Выбор наибольшего из этих I3 чисел определяет оптимальное значение признака тон/шум и периода ОТ.

Итак, предлагаемый метод применим и в случае, когда траекторию ОТ можно определять после завершения ввода всей реализации, и в случае оперативного оценивания.

В первом случае ошибки в значениях ОТ практически исключаются. Во втором случае вероятность ошибки после 4–5 вокализованных интервалов также довольно мала.

УПРОЩЕННАЯ СХЕМА: МОДИФИКАЦИЯ КОРРЕЛЯЦИОННОГО МЕТОДА

Ниже предлагается упрощенный вариант, являющийся фактически усовершенствованной модификацией корреляционного метода, поскольку в нем принятие решений осуществляется на основании сумм коэффициентов автокорреляции на траектории ОТ.

Сначала для каждого состояния модели $S_n = (s_n, i_n, v_n)$ определяется вес:

$$d(R_n,S_n) = \begin{cases} R_n(s_n), & \text{если } v_n = 1, \\ A - \max\limits_{s} R_n(s), & \text{если } v_n = 0. \end{cases}$$

где A – эмпирическая константа. Затем для каждой пары состояний, между которыми возможен переход определяется величина:

$$g(R_n,S_{n-1},S_n) = \begin{cases} d(R_n,S_n), & \text{если } s_n = s_{n-1}, \\ d(R_n,S_n) - D, & \text{если } s_n = s_{n-1} \pm 1, \end{cases}$$

где D – также эмпирическая константа.

Теперь ставится задача поиска такой последовательности состояний S_n, $1 \leqslant n \leqslant N$, связанных допустимыми переходами, которая имела бы наибольшую сумму:

$$F = \sum_{n=1}^{N} g(R_n,S_{n-1},S_n).$$

Эта задача решается алгоритмом (2).

Описанный упрощенный вариант проверялся на тестовом материале из 9 фраз, произнесенных двумя дикторами-женщинами и одним диктором-мужчиной; частота дискретизации I6 кГц, порядок обратного фильтра I4, частота среза НЧ-фильтра I,25 кГц, длина интервала анализа 30 мс, шаг – 20 мс. Общий объем речевого сигнала около 800 интервалов анализа. Эксперименты показали высокую надежность алгоритма: не было ни одного ошибочного значения периода ОТ. Были только некоторые сомнения относительно признака тон/шум на стыках вокализованных и невокализованных участков.

ЗАКЛЮЧЕНИЕ

Предварительные эксперименты говорят о перспективности применения марковской модели для надежного выделения ОТ. При этом роль модели будет тем больше, чем менее информативным является используемое первичное описание, что имеет место, например, в случае зашумленной речи.