УДК 519.254

АНАЛИЗ СКРЫТЫХ ТРАЕКТОРНЫХ МОДЕЛЕЙ РЕЗОНАТОРОВ РЕЧЕВОГО ТРАКТА ДЛЯ СИСТЕМ РАСПОЗНАВАНИЯ ФОНЕМ

***Леднов Д.А.,** к. т. н., старший научный сотрудник, научный консультант научно-технического департамента ООО «Стэл – Компьютерные Системы», г. Москва, e-mail: lednov@stel.ru*

Ключевые слова: распознавание фонем, речевой тракт, траекторная модель, резонатор, КИХ-фильтр, линейное предсказание, оптимизация.

Приводится перевод и анализ оригинальных работ Li Deng и его коллег (компания Microsoft), появившихся в период с 2000 по 2010 год в области фонетического распознавания речи. Основное направление этих публикаций связано с разработкой модели скрытых траекторий параметров резонаторов вокального тракта. В ходе разработки было показано, как функционально зависит динамика коэффициентов линейного предсказания от параметров резонаторов вокального тракта, которые предварительно сглаживаются КИХ-фильтром. Затем, для этой зависимости введена статистическая модель, для которой поставлена и решена оптимизационная задача. Автор настоящей работы дополнил развитую модель уравнением непрерывности, которое позволяет определить характеристики КИХ-фильтров для каждого фонетического состояния и ввел альтернативную оптимизационную схему, позволяющую определять параметры статистической модели.

Введение

Согласно классической схеме систем распознавания слитной речи [1], система распознавания фонем является частью системы распознавания речи и, в основном, именно точность распознавания фонем определяет точность системы распознавания в целом.

В основе большинства современных систем распознавания фонем лежит скрытая модель Маркова (СММ), которая описывает вероятности следования состояний, описывающих фонемы, друг за другом. Для описания состояний фонем используются смеси нормальных плотностей распределения (в литературе их чаще называют гауссовыми смесями) векторов наблюдений.

В соответствии с работой [2], средняя точность системы распознавания фонем (по всему множеству фонем английского языка), которая достигнута на современном уровне технологий в условиях не зашумленного сигнала, при частоте оцифровки 8кГц и при использовании СММ, составляет 69.1%. Эта величина не является высокой, поэтому актуальной является задача поиска новых математических моделей для систем распознавания фонем.

В период с 2000 по 2010 год появилась серия публикаций *Li Deng* и его коллег (компания Microsoft) [4-8], в которых описывается новый подход к задаче распознавания фонем, основанный на фундаментальном свойстве артикуляционных органов – их инерции.

Инерция артикуляционных органов отражается в речи двумя свойствами:

– гладкостью изменения ее параметров во времени в процессе произнесения вокализованных звуков;

– коартикуляцией [3].

Гладкость изменений параметров речи проявляется в том, что амплитуды и частоты гармоник линейчатого спектра, порожденного вокализованным звуком, не изменяются скачкообразно. Коартикуляция заключается во влиянии целевых артикуляций соседних звуков друг на друга.

Авторы [4-8] назвали построенный способ распознавания фонем на основе инерции артикуляционных органов «скрытыми траекторными моделями» (СТМ) резонаторов речевого тракта (РРТ). Здесь термин «скрытость», как и в модели Маркова, заключается в том, что при регистрации речи мы можем измерить только физические параметры речи (вектора наблюдений), но информационная ее составляющая (последовательность фонем), которую хочет передать говорящий, является для нас скрытой и определяется с некоторой вероятностью в результате анализа.

Работа над пониманием названных выше публикаций выявила, с одной стороны, допущения, сделанные авторами, которые искажают первоначальную идею модели, в результате чего траектории параметров РРТ становятся разрывными. С другой стороны, авторы не указали способ определения начального приближения плотности вероятности параметров модели.

Настоящая работа ставит перед собой цель – изложение основных теоретических положений скрытых траекторных моделей резонаторов речевого тракта и устранение в ней указанных ранее недостатков.

Генерация скрытых траекторий параметров резонаторов речевого тракта

Схематически идею СТМ параметров РРТ можно представить, как показано на рис. 1. Это схема состоит из двух частей: модели генерации речи; модели восприятия и обработки речи. В стадию обработки речи включены два процесса – обучение (оценка параметров модели) и распознавание (принятие решение о звучащей фонеме или их последовательности).

В соответствии с рис. 1, блок генерации символов преобразует целевые семантические категории в последовательность слов, затем последовательность слов - в последовательность фонем, которые поступают на вход блока управления речевым трактом (РТ).

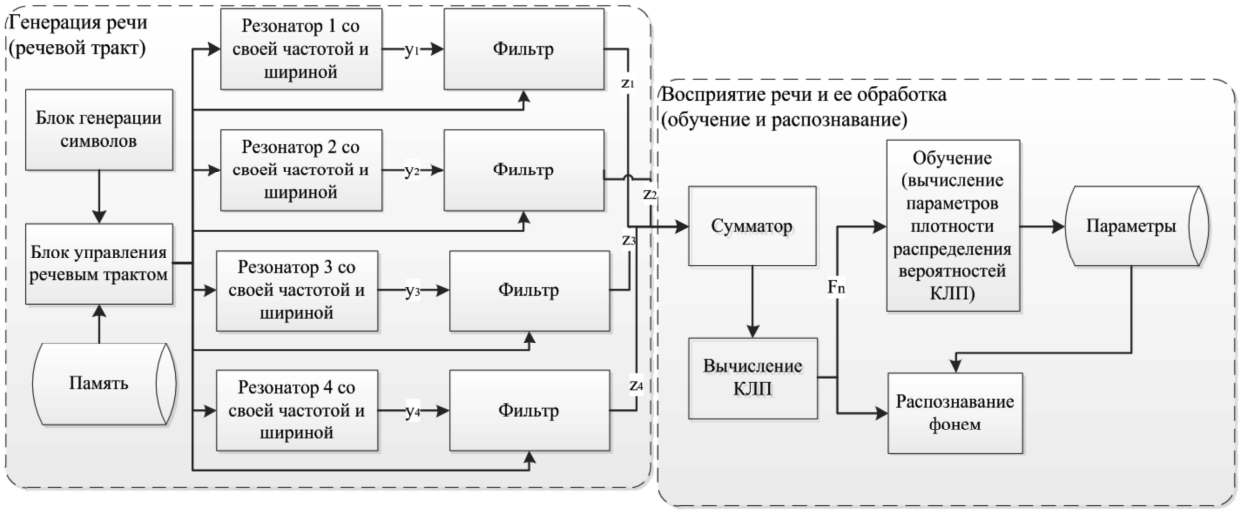

Рис. 1. Схематическое представление идеи СТМ для РРТ

В блоке управления РТ каждой поступающей фонеме ставится в соответствие набор целевых параметров РТ, определяющих его форму (целевые параметры РТ извлекаются из памяти). Блок управления РТ вызывает динамику речевого тракта, целью которой является форма РТ, определяющая звучание фонемы. Заметим, что РТ может и не достигнуть заданной формы, это зависит от темпа смены фонем в их последовательности, поступающей на вход управляющего блока.

В СТМ речевой тракт представлен четырьмя резонаторами, которые характеризуются частотой и шириной резонанса, а также сглаживающими КИХ-фильтрами одинаковыми для всех резонаторов. При смене одной фонемы другой блок управления изменяет целевые параметры резонаторов и импульсную характеристику сглаживающих фильтров. Источником возбуждения резонаторов является широкополосный белый шум.

С одной стороны, можно считать недостатком то, что модель пренебрегает фактом существования голосового источника, порождающего гармонические линейчатые спектры, но, с другой стороны, это предположение позволяет ввести универсальное описание акустических свойств фонем, независимое от их типа (вокализованные, невокализованные) на основе плотности распределения энергии в спектре акустического сигнала.

Модель восприятия и обработки речи предполагает, что параметризация речевого сигнала происходит с помощью известной процедуры вычисления коэффициентов линейного предсказания (КЛП). Ниже будет показана явная зависимость КЛП от параметров резонаторов и, как следствие, зависимость вектора наблюдения от параметров резонаторов в условиях сглаживания КИХ-фильтрами.

И далее, в соответствии с классической схемой систем распознавания [1], в процессе обучения происходит вычисление параметров (математического ожидания и ковариационной матрицы) плотностей распределения вероятности векторов наблюдений (плотности вероятности предполагаются нормальными), после чего в процессе распознавания происходит вычисление наиболее вероятной цепочки фонем на основе входной последовательности векторов наблюдений.

Формализуем описанный процесс генерации речи. С этой цель составим фонетически зависимый целевой вектор T_s (target) из параметров резонаторов, где s–индекс фонемы.

Условное распределение динамики текущих параметров y резонаторов, при заданном целевом векторе, определенном звучащей фонемой s, предположим нормальным:

$$p(y|s) = G(y;\, \mu_{T_s} \sum\nolimits_{T_s}),$$

где $\mu_{T_s}, \sum_{T_s}$ – математическое ожидание и ковариационная матрица параметров резонаторов.

Скрытый случайный процесс $z(t)$ образуется путем фильтрации целевого процесса $y_s(t)$. Эта фильтрация выполняется с помощью фильтра с конечной импульсной характеристикой вида:

$$h_s(t) = \begin{cases} g_s \gamma_{s(t)}^{-t}, & -D < t < 0; \\ g_s, & t = 0; \\ g_s \gamma_{s(t)}^{t}, & 0 < t < D, \end{cases}$$

где g_s – нормализующая константа, $\gamma_{s(t)}^{t}$ – параметр шкалы (лежит в интервале [0,1]).

Скрытый случайный фонетически зависимый процесс $z_s(t)$ можно представить в виде свертки целевого процесса и импульсной характеристики:

$$z_s(t) = h_s(t) * y_s(t) = \sum_{t=k-D}^{t=k+D} g_s \gamma_{s(t)}^{|t-k|} y_s(t),$$

где $y_s(t)$ текущее значение параметра резонатора при заданной фонеме s.

Линейная зависимость между скрытым процессом $z_s(t)$ и целевым процессом $y_s(t)$ приводят к линейным зависимостям между математическим ожиданиями целевого вектора и скрытого процесса:

$$\mu_s(t) = \sum_{t=k-D}^{t=k+D} g_s \gamma_{s(t)}^{|t-k|} \mu_{Ts(t)} = \mathrm{a}_k \mu_T.$$

Каждая f-ая компонента вектора $\mu_s(t)$ может быть представлена в виде:

$$\mu_{z(t)}(f) = \sum_{l=1}^{\Phi} a_t(l)\mu_T(l, f), \quad (1)$$

где Φ – количество фонетических состояний и f = 1,..,8 для четырех частот и соответствующих им полос пропускания.

Сходным образом для ковариационной матрицы справедливо

$$\sum_{z(k)} = \sum_{t=k-D}^{t=k+D} g_s^2 \gamma_{s(t)}^{2|t-k|} \sum_{s(t)}.$$

Аппроксимируем ковариационную матрицу диагональной матрицей

$$\sigma_{z(k)}^2 = \vartheta_k \sigma_T^2$$

или в покомпонентной форме

$$\sigma_{z(t)}^2(f) = \sum_{l=1}^{\Phi} \vartheta_t(l)\sigma_T^2(l, f). \quad (2)$$

В работах [4-6] нет прямых указаний о способе вычисления коэффициентов a_k и ϑ_k. Авторы пишут, что нет краткой математической формы, в которой можно записать выражение для них.

Рассмотрим вопрос, - а можно ли в действительности получить такую математическую форму? Пусть k это момент времени, в который помещена середина «окна» импульсной характеристики наблюдения размером 2D+1. Необходимо вычислить влияние сегмента речи, заключенного в интервале времени d (см. рис. 2), на систему в текущий момент времени k. Можно выделить три различных случая положения границ d звучания этого сегмента речи (L-левая граница, R-правая граница) относительно середины окна наблюдения k, которые показаны на рис. 2.

Для случая, показанного на рис. 2 a, где $k - D \le L < k$ и $k \le R \le k + D$, справедливо

$$a_k = g\left(\sum_{t=L}^{k} \gamma^{|t-k|} + \sum_{t=k}^{R} \gamma^{|t-k|}\right) = \frac{g}{1-\gamma}\left(2 - \gamma^{k-L+1} - \gamma^{R-k+1}\right). \quad (3.1)$$

Для случая, показанного на рис. 2 b, где $k - D \le L < k$ и $L \le R \le k$, справедливо

$$a_k = g\gamma^{R-k}\frac{1-\gamma^{R-L+1}}{1-\gamma}. \quad (3.2)$$

Для случая, показанного на рис. 2 c, где $k \le L < k$ и $k \le R \le k + D$, справедливо

$$a_k = g\gamma^{k-L}\frac{1-\gamma^{R-L+1}}{1-\gamma}. \quad (3.3)$$

Таким образом, для каждого типа положения границ фонемы относительно центра «окна» импульсной характеристики можно найти простую математическую форму записи коэффициентов в (1) и аналогично в (2), причем

$$\sum_{t=k-D}^{t=k+D} a_t(l) = 1 \quad \forall l. \quad (4)$$

Откуда можно вычислить значение коэффициента $g = (1-\gamma)/2(1-\gamma^{(D+1)})$.

Для определения параметра γ можно использовать условие непрерывности, физический смысл которого состоит в том, что изменения значения траектории должны быть малы при малом влиянии фонемы. Пусть в момент времени $k - D + 1$ одна фонема меняет другую и на протяжении времени от $k - D + 1$ до $k + D$ звучит некоторая определенная последовательность фонем. Затем «окно» импульсной характеристики смещается на шаг вправо, при этом крайняя левая фонема исчезает из поля «окна», а остальная последовательность фонем сохраняется. В этом случае можно записать условие непрерывности

$$g_1\gamma_1^{1-D}(1+\gamma_1)\mu_{1T} + \sum_{i=2}^{N} g_i^{(k)}\mu_{iT}\sum_{t=L_i}^{t=R_i}\gamma_i^{|t-k|} \approx \approx \sum_{i=2}^{n} g_i^{(k+1)}\mu_{iT}\sum_{t=L_i}^{t=R_i}\gamma_i^{|t-k-1|}. \quad (5)$$

где N+1 – количество фонем, которые укладываются в окне длительностью $2D + 1$.

Обратим внимание на то, что в работе [4] для всех типов акустических данных используются однотипные КИХ-фильтры, а в работе [5] предполагается, что тип КИХ-фильтра не зависит от фонемы, т.е. $\gamma = \gamma_s$ для всех s. В дальнейшем мы используем (5) для определения параметров модели фильтров, метод получения которых не был описан в анализируемых работах.

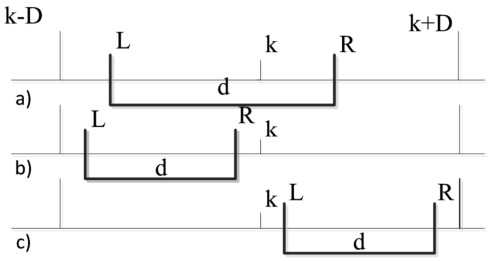

Рис.2. Различные положения границ звучания фонемы относительно центра границ импульсной характеристики

Генерация акустических данных

В работе [6] получена аналитическая форма коэффициентов линейного предсказания (КЛП) для случая резонансной модели речеобразования.

Рассмотрим комплексные корни передаточной функции речевого тракта:

$$z_m = \exp\left\{-\pi\frac{b_m}{f_s} + j2\pi\frac{f_m}{f_s}\right\}, \quad (6)$$

где f_s – частота оцифровки речевого сигнала, $b_p(t)$ – ширина резонанса, $f_p(t)$ – частота резонанса.

Передаточную функцию речевого тракта с K полюсами можно записать в виде

$$H(t) = G\prod_{k=1}^{K}\frac{1}{(1-z_k z^{-1})(1-z_k^* z^{-1})},$$

или в логарифмической форме

$$\log H(z) =$$

$$= \log G - \sum_{k=1}^{K}\log(1-z_k z^{-1}) - \sum_{k=1}^{K}\log(1-z_k^* z^{-1}).$$

Если использовать известное разложение логарифма

$$\log(1-x) = -\sum_{i=1}^{\infty}\frac{x^i}{i},$$

то последнее выражение можно представить в форме

$$\log H(z) = \log G - \sum_{k=1}^{K}\sum_{i=1}^{\infty}\frac{z_k^i z^{-i}}{i} - \sum_{k=1}^{K}\sum_{i=1}^{\infty}\frac{z_k^* z^{-i}}{i} =$$

$$= \log G + \sum_{i=1}^{\infty}\left[\sum_{k=1}^{K}\frac{z_k^i + z_k^{*i}}{i}\right]z^{-i} = c_0 + \sum_{i=1}^{\infty}c_i z^{-i}.$$

где

$$c_0 = \log G,$$

$$c_i = \sum_{k=1}^{K}\frac{z_k^i + z_k^{*i}}{i}.$$

Если подставить (6) в формулу для коэффициентов c_i, то несложно получить

$$c_n = \frac{2}{n}\sum_{k=1}^{K}\exp\left\{-\pi n\frac{b_k}{f_s}\right\}\cos\left(2\pi n\frac{f_k}{f_s}\right).$$

Из последней формулы непосредственно следует, что для n-компонентного значения вектора наблюдений, порожденного скрытой векторной функцией $z(t)$, справедливо

$$F_n = (z(t) = \{f_p, b_p\}_p) =$$

$$= \frac{2}{n}\sum_{p=1}^{P}\exp\left\{-\frac{\pi n b_p(t)}{f_s}\right\}\cos\left(\frac{2\pi n f_p(t)}{f_s}\right),$$

P – количество резонансов (P = 4), f_s – частота оцифровки.

Остаток от аппроксимации наблюдения можно записать в виде

$$r_s(t) = o(t) - F(\{z_s(t) = \{f_p, b_p\}_p).$$

Предположим, что плотность распределения величины остаточного вектора - это случайная величина, заданная нормальным распределением

$$p(r_s(t) \mid z(t), s) = G(r_s(t);\, \mu_r, \Sigma_r), \qquad (7)$$

s –индекс фонемы.

Из последнего предположения следует, что условное распределение наблюдения можно записать в виде:

$$p(o(t) \mid z(t), s) = G(o(t);\, F(z_s(t)) + \mu_r, \Sigma_r), \qquad (8)$$

и представить вектор наблюдения в форме:

$$o(t) = F(z_s(t)) + \mu_r + w_s(t),$$

где $w_s(t)$ – случайная величина с нормальным распределением $G(w_s(t), 0, \Sigma_r)$.

Линеаризация коэффициентов линейного предсказания

Нелинейная функция $F(z_s(t))$ сложна для дальнейших вычислений, поэтому проведем ее линеаризацию. С помощью разложения в ряд Тейлора с точностью до первого порядка представим функцию $F(z_s(t))$ в виде:

$$F(z(t)) = F(z_0(t)) + F'(z_0(t))(z(t) - z_0(t)),$$

где компоненты матрицы Якоби могут быть представлены в форме:

$$F_n'(f_p(t)) = \frac{-4\pi}{f_0}\exp\left\{-\frac{\pi n b_p(t)}{f_0}\right\}\sin\left(\frac{2\pi n f_p(t)}{f_0}\right),$$

$$F_n'(b_p(t)) = \frac{-2\pi}{f_0}\exp\left\{-\frac{\pi n b_p(t)}{f_0}\right\}\cos\left(\frac{2\pi n f_p(t)}{f_0}\right).$$

Подстановка разложения в ряд Тейлора приведет к новой форме для условной вероятности (8)

$$p(o(t) \mid z(t), s) = G(o(t);\, F(z_0(t)) +$$

$$+F'(z_0(t))(z(t) - z_0(t)) + \mu_r, \Sigma_r),$$

где можно ввести обозначение математического ожидания вектора наблюдения

$$\mu_0(t) = F(z_0(t)) + F'(z_0(t))(z(t) - z_0(t)) + \mu_r(t),$$

и в сокращенном виде записать

$$p(o(t) \mid z(t), s = G(o(t);\, \mu_0(t), \Sigma_r).$$

Вычисление вероятности параметров РРТ

Рассмотрим условную вероятность возникновения наблюдения при звучании известной фонемы

$$p(o(t) \mid s = \int p(o(t) \mid z(t), s)\, p(z(t) \mid s) dz \sim$$

$$\sim \int G(o(t); \mu_{o(t)}, \Sigma_{r(t)}, s) G(z(t)\,; \mu_{z(t)}, \Sigma_{z(t)}, s)\, dz =$$

$$= G(o(t); \mu_{o(t)}, \Sigma_{o(t)}, s), \qquad (9)$$

где введено новое обозначение

$$\Sigma_{0(t,s)}(t) = \Sigma_{r(t,s)}(t) + F'(z_0(t))\Sigma_{z(t)}[F'(z_0(t))]^T.$$

Раскроем введенные обозначения математического ожидания и ковариационной матрицы наблюдения, используя выражения (1), (2), и для сокращения записи выпишем отдельно числитель h и знаменатель v экспоненты для j-ой компоненты вектора наблюдения:

$$h = \left(o_t(j) - F_j(z_o(t)) - \mu_r(j) -\right.$$

$$\left. - F_j'(z_0(t))\left(\sum_{l=1}^{\Phi}a_t(l)\mu_T(l,f) - z_o(t)\right)\right)^2,$$

$$v = \sigma^2_{r(t,s)}(j) + \\ + \sum_{k=1}^{P} F'_{jk}(z_o(t)) F'^{T}_{kj}(z_o(t)) \sum_{l=1}^{\Phi} \vartheta_t(l) \sigma^2_T(l,k) \qquad (11)$$

Из приведенных формул для числителя и знаменателя видно, что неизвестными параметрами модели являются математические ожидания $\mu_T(l,f)$ и дисперсии $\sigma^2_T(l,k)$ целевых векторов фонем, математические ожидания $\mu_{r(t)}(j)$ и дисперсии $\sigma^2_{r(t,s)}(j)$ остаточного вектора и коэффициенты свертки математических ожиданий $a_t(l)$ и коэффициенты свертки дисперсий $\vartheta_t(l)$ целевых векторов. Проведем оценку перечисленных параметров.

Оценка параметров для СТМ

Оценка параметров модели происходят на размеченной речевой базе данных, принципы разметки данных описаны в работе [7].

В оригинальных статьях [4, 8] параметры распределения (9) были разделены на две группы: параметры распределения кепстрального остатка и значения компонент целевых векторов. Метод оценки величин γ_s (или даже величины $\gamma = \gamma_s \ \forall_s$, в соответствии с [5]) авторы не приводят. Здесь сначала приводится оценка для приведенных групп параметров, а затем эта схема будет дополнена.

Оценка параметров распределения кепстрального остатка

Для определения параметров распределения кепстрального остатка найдем максимум функции логарифмического правдоподобия. Для математического ожидания справедливо уравнение:

$$\frac{\partial \log \prod_{t=1}^{K_s} p(o(t) \mid s)}{\partial \mu_{r(s)}} = 0,$$

где K_s - количество дискретных отсчетов времени, в которых наблюдалась фонема *s*.

Выполняя дифференцирование последнего уравнения для математического ожидания остатка, получим:

$$\mu_{r(s)} = \qquad (12)$$
$$= \frac{\sum_{t=1}^{K} [o(t) - F'(z_0(t)) \mu_{z(t)} - F(z_0(t)) + F'(z_0(t)) z_0(t)]}{K_s}$$

Аналогично, для диагональной ковариационной матрицы справедливо уравнение

$$\frac{\partial \log \prod_{t=1}^{K_s} p(o(t) \mid s)}{\partial \sigma^2_{r(s)}} = 0,$$

которое приводит к выражению

$$\sum_{t=1}^{K_s} \frac{\sigma^2_{r(s)} + q(t) - (o(t) - \mu_{0(s)})^2}{(\sigma^2_{r(s)} + q(t))^2} = 0, \qquad (13)$$

где введено обозначение

$$q(t) = diag\{F'(z_0(t)) \sum_{z(t)} [F'(z_0(t))]^T\}. \qquad (14)$$

Уравнение (13) можно разрешить тремя способами:
1) предположить, что $q(t)$ не зависит от времени;
2) методом градиентного спуска;
3) методом ограниченного градиентного спуска.

Оценка параметров основного распределения

Подмножество параметров СТМ состоит из: 1) средних векторов $\mu_{0(s,t)}(j)$; 2) диагональных элементов ковариационной матрицы $\sigma^2_{0(t,s)}(j)$.

1) Средние вектора

Ранее была предположена диагональная форма ковариационной матрицы.

Запишем многомерное распределение

$$p(o(t) \mid s(t)) = \\ = \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma^2_{o(t,s)}(j)}} \exp\left\{ \frac{(o_t(j) - \mu_{o(s,t)}(j))^2}{\sigma^2_{o(t,s)}(j)} \right\},$$

$\sigma^2_{0(t,s)}(j)$ - *j*-ая компонента наблюдения во фрейме в момент времени *t*, полученный от фонемы *s*.

Функция логарифмического правдоподобия относительно среднего вектора $\mu_{0(s,t)}(j)$ может быть записана в виде

$$P = \sum_{t=1}^{K(s)} \sum_{j=1}^{J} \left\{ \frac{(o_t(j) - \mu_{o(s,t)}(j))^2}{\sigma^2_{o(t,s)}(j)} \right\} = \qquad (15)$$
$$= \sum_{t=1}^{K(s)} \sum_{j=1}^{J} \left\{ \frac{\sum_f F'(z_0(t), j, f) \sum_l a_t(l) \mu_T(l,f) - d_t(j))^2}{\sigma^2_{o(t,s)}(j)} \right\},$$

где l и f – индексы фонем и PPT компонент соответственно, и

$$d_t(j) = o_t(j) - F(z_0(t), j) + \\ + \sum_f F'(z_0(t), j, f) z_0(t, f) - \mu_{r(s,t)}(j).$$

Если выполнить дифференцирование (15)

$$\frac{\partial P}{\partial \mu_T(l_0, f_0)} = 0$$

и перегруппировать полученное выражение так, чтобы слагаемые, содержащие $\mu_T(l,f)$ были слева, а все прочие элементы справа, то получим уравнение

$$\sum_f \sum_l A(l, f, l_0, f_0) \mu_T(l,f) = \\ = \sum_t \left\{ \sum_j \frac{F'(z_0(t), j, f_0) d_t(j)}{\sigma^2_{o(t,s)}(j)} \right\} a_t(l_0), \qquad (16)$$

где $f_0 = 1,2,...,8$ для каждой размерности $l_0 = 1,2,...,58$ для каждого фонетического элемента

$$A(l,f,l_0,f_0) =$$

$$= \sum_{t,j} \frac{F'(z_0(t),j,f_0)F'(z_0(t),j,f_0)}{\sigma^2_{o(t,s)}(j)} a_t(l)a_t(l_0).$$

2) Матрица ковариаций

Для матрицы ковариации введем функцию правдоподобия

$$L \sim -\sum_{t=1}^{K}\sum_{j=1}^{J}\left\{\frac{(o_t(j)-\mu_{0(s,t)}(j))^2}{\sigma^2_{o(t,s)}(j)+q(t,j)} - \log(\sigma^2_{z(t,s)}(j)+q(t,j))\right\}$$

где $q(t,j)$ – элемент вектора, определенный в (14).

Используя метод градиентного спуска, получим

$$\frac{\partial L}{\partial \sigma^2_T(l,f)} =$$

$$= \sum_{t=1}^{K}\sum_{j=1}^{J}\left\{\frac{(o_t(j)-\mu_{0(s,t)}(j))^2 F'^2_{jf}\vartheta_t(l)}{(\sigma^2_{r(s)}(j)+q(t,j))^2} - \frac{F'^2_{jf}\vartheta_t(l)}{\sigma^2_{r(s)}(j)+q(t,j)})\right\}$$

$$\sigma^2_T(l,f) \leftarrow \sigma^2_T(l,f) + \frac{\partial L}{\partial \sigma^2_T(l,f)}. \qquad (17)$$

Резюмируем приведенные вычисления. Схема поиска параметров, предложенная в работах [4-6], состоит в рекуррентном решении уравнений (12), (13) и (16), (17) с помощью классических методов градиентного спуска. К сожалению, в работах не обсуждается проблема выбора начального приближения для этих рекуррентных вычислений.

Альтернативная схема оценки параметров

Прежде всего, решим вопрос относительно начального приближения. Для вычисления начальных приближений средних частот и ширины полосы резонансов выберем всевозможные спектры, принадлежащие определенной фонеме s из обучающей речевой базы. Частотный диапазон спектров разобьём на перекрывающиеся участки с границами, указанными в работе [6] (табл. 1).

Таблица 1. Список спектральных границ для вычисления начального приближения

Номер резонанса	(rb) Правая граница (Гц)	(lb) Левая граница(Гц)
1.	200	900
2.	600	2800
3.	1400	3800
4.	1700	5000

Шаг 1. В качестве частоты i-го резонанса в начальном приближении используем среднее значение частоты в i-ом спектральном диапазоне

$$f_i^{(0)} = \frac{1}{K_s}\sum_{t=1}^{K_s}\frac{\int_{lb_i}^{rb_i} S_t(\omega)\omega d\omega}{\int_{lb_i}^{rb_i} S_t(\omega)d\omega},$$

а в качестве ширины полосы частот i-го резонанса среднеквадратическое отклонение

$$b_i^{(0)} = \sqrt{\frac{1}{K_s}\sum_{t=1}^{K_s}\frac{\int_{lb_i}^{rb_i} S_t(\omega)\left(\omega - f_i^{(0)}\right)^2 d\omega}{\int_{lb_i}^{rb_i} S_t(\omega)d\omega}},$$

где $S_t(\omega)$ спектр фонемы, наблюдаемый в момент времени t. Таким образом, построены начальные приближения целевых векторов фонем $\mu_s^{(0)} = \left(\left\{f_i^0\right\}_s, \left\{b_i^0\right\}_s\right)^T$.

Шаг 2. Рассмотрим последовательности фонем, которые удовлетворяют условиям построения уравнений (5). Из всевозможных последовательностей будем выбирать такие подпоследовательности, на основе которых можно строить разрешимые нелинейные системы уравнений (5). Каждая из таких систем будет определять множество решений $\{\gamma\}$ для фонем, входящих в подпоследовательность. Очевидно, что одна и та же фонема может входить в различные подпоследовательности и для нее может быть получено множество решений. Для полученных решений построим гистограмму их встречаемости и выберем наиболее часто встречающиеся решения. Так поступим с множествами решений для каждой фонемы.

Шаг 3. Предположим, что

$$z_0(t) = \sum_{l=1}^{\Phi} a_t^{(0)}(l)\mu_T^{(0)}(l,f).$$

Подстановка этого равенства в (10) и (11) и дифференцирование правдоподобия по параметру математического ожидания кепстрального остатка и его дисперсии дает нулевое приближение этих параметров:

$$\mu^{(0)}_{r(t)}(j) = \frac{1}{K_s}\sum_{t=1}^{K_s}\left[o_t(j) - F_j(z_0(t))\right];$$

$$\left(\sigma^{(0)}_{r(t,s)}(j)\right)^2 = \left(o_t(j) - F_j(z_0(t)) - \mu^{(0)}_{r(t)}(j)\right)^2 -$$

$$-\sum_{k=1}^{P} F'_{jk}(z_0(t))F'^T_{jk}(z_0(t))\sum_{l=1}^{\Phi}\vartheta_t^{(0)}(l)\left(\sigma_T^{(0)}(l,k)\right)^2.$$

Шаг 4. Расчет первого приближения параметров резонансов вокального тракта происходит на основании уравнений (16) и (17).

Шаг 5. На основании нового приближения параметров резонансов вокального тракта возвращаемся к шагу 2 и решаем системы уравнений (5) с новыми параметрами.

Необходимо отметить, что в последнем слагаемом (10) в выражении в скобках, свертка $\sum_{l=1}^{\Phi} a_t(l)\mu_T(l,f)$ от $z_0(t)$ отличается лишь тем, что она вычисляется на основании параметров, вычисленных на одну итерацию выше.

Экспериментальные результаты

Экспериментальные результаты были получены с помощью корпуса речи TIMIT (частота оцифровки 16кГц). Для обучения были выделены 538 предложений [9]. Эти предложения были фонетически размечены и найдены формантные траектории, на основе которых проведена оценка параметров модели (12), (14), (16) и (17). На остальной части корпуса речи TIMIT было проведено сравнение точностей распознавания фонем, полученных с помощью скрытой модели Маркова и скрытой траекторной модели. Скрытая модель Маркова была реализована в следующей форме: в качестве вектора признаков использовались кепстральные коэффициенты линейного

предсказания; модель фонем представлена трифонами, включающими в себя три состояния; использовалась биграммная модель языка. Точности, показанные этими системами, приведены в табл. 2.

Таблица 2. Сравнение точностей распознавания фонем скрытой модели Маркова и скрытой траекторной модели

Тип фонем	Сонорные (гласные, звонкие согласные, нозальные)	Взрывные	Фрикативные	Смычные
Количество фонем	3814	889	1252	1578
СММ	64.05	72.10	75.74	88.72
СТМ	72.42	76.27	75.74	90.94

Из табл. 2 видно, что по сравнению со скрытой моделью Маркова модель СТМ дает существенное преимущество в точности.

Заключение

Как известно, речь обладает существенной вариативностью, и поведение резонаторов речевого тракта при реализации той или иной фонемы может изменяться от одного говорящего к другому [11]. Если предположить, что такие различные типы поведения речевого тракта можно объединять в группы и каждая группа обладает сходными характеристиками и динамикой речевого тракта, то можно заключить, что распределение плотности вероятности (7) справедливо только в рамках одной группы. Тогда рассуждения, которые приводят к выражению (9), остаются справедливыми, если допустить, что нам априори известна принадлежность голоса говорящего к той или иной группе. Если же такой информации нет, то (9) можно записать в форме:

$$p(o(t) \mid s) = \sum_{i=1}^{m(s)} p_i(s) G(o(t); \mu_{o(t),i}; \sum_{o(t),i} s), \quad (17)$$

где m – количество групп голосов для каждой фонемы, $p_i(s)$ вероятность появления i-ой группы фонемы s. С точки зрения автора, введенная гауссова смесь (17) способна в полном объеме придать системе распознавания фонем свойство дикторонезависимости и описать вариативность речи говорящих.

Таким образом, в работе описана модель скрытых траекторий параметров резонаторов речевого тракта, введены поправки к исходной модели, а также предложен путь развития модели на основе описания фонем - как гауссовых смесей скрытых траекторий.

Поиск информации о продолжении работы в направлении развития СТМ для РРТ после 2010 года не принес результатов.

Литература

1. Jelinek F. Continuous speech recognition by statistical methods.// In: Proceedings of the IEEE, 1976, v. 64, № 4, p. 532–556.

2. Thomas, S.; Ganapathy, S. &Hermansky, H. (2008), Spectro-Temporal Features for Automatic Speech Recognition using Linear Prediction in Spectral Domain//in 'EUSIPCO'

3. Кодзасов С.В., Кривнова О.Ф. Общая фонетика: Учебник. М.: Рос.гос. гуманит. ун-т, 2001. 592с.

4. Li Deng, Dong Yu and Alex Acero, Structured Speech Modeling// IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 5, Sep. 2006

5. Li Deng and et al., Acoustic Models with Structured Hidden Dynamics with Integration Over Many Possible Hidden Trajectories// US Patent No. 7,565,284 B2, Date of Patent Jul. 21, 2009

6. Li Deng, Alex Aceroand IssamBazzi, Tracking Vocal Tract Resonances Using a Quantized Nonlinear Function Embedded in a Temporal Constraint // IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 2, March 2006

7. Leo Jingyu Lee, Hidden Dynamic Models for Speech Processing Applications// A thesis presented to the University of Waterloo Waterloo, Ontario, Canada, 2004

8. Li Deng, Dong Yu, Xiaolong Li and Alex Acero, A long-contextual-span model of resonance dynamics for speech recognition: parameter learning and recognizer evaluation// IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU ’05), Nov 27 – Dec 1, Cancun, Mexico, 2005.

9. Li Deng, Xiaodong Cui, Robert Pruvenok, Jonathan Huang, SafiyyMomen, Yanyi Chenand AbeerAlwanA Databaseof Vocal Tract Resonance Trajectoriesfor Researchin Speech Processing//Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), May, 2006

10. Потапова Р. К., Потапов В. В. Речевая коммуникация: От звука к высказыванию − М.: Языки славянских культур, 2012. - 464 с.

ANALYSIS OF HIDDEN TRAJECTORY MODELS (HTM) OF VOCAL TRACT RESONANCE (VTR) FOR PHONEMES RECOGNITION SYSTEMS

Lednov D.A.

In this paper there is given the translation to Russian and drawn the analysis of original In this paper there is given the translation to Russian and drawn the analysis of original works of Li Deng and his colleagues devoted to phoneme recognition publishedfrom 2000 to 2009. The main direction of these publications is the development of hidden trajectory models of vocal tract resonance. There has been illustrated the functional dependence between linear prediction coefficients and resonators parameters which are preliminarily smoothed by the FIR-filter. Optimization of a statistical model introduced for this functional dependence is performed. In this paper the above model is completed with continuity equation which enables to determine the FIR-filter features and an alternative optimization scheme is introduced which allows to estimate the statistical model parameters.

The main direction of these publications is the development of hidden trajectory models of vocal tract resonance. There has been illustrated the functional dependence between linear prediction coefficients and resonators parameters which are preliminarily smoothed by the FIR-filter. Optimization of a statistical model introduced for this functional dependence is performed. In this paper the above model is completed with continuity equation which enables to determine the FIR-filter features and an alternative optimization scheme is introduced which allows to estimate the statistical model parameters.