Информационные процессы, Том 24, № 2, 2024, стр. 119–135.

= МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ =

Об одной модели поверхности языка в средне-сагиттальном сечении

И.С. Макаров

ООО "БиометрикЛабс Россия, Москва, im@biometriclabs.ru

Поступила в редколлегию 01.05.2024 г. Принята 17.07.2024 г.

Аннотация—Построена математическая модель языка в средне-сагиттальном сечении. В рамках модели профиль языка описывается гибким упругим стержнем, для которого численно решается уравнение упругой линии с заданными граничными условиями и распределенными внешними силами. База контуров языка, собранная из решений этого уравнения, подвергается кластеризации методом К-средних на шестнадцать классов. Результирующий контур языка представляет собой линейную комбинацию центроидов этих классов, при этом коэффициенты при центроидах удовлетворяют ряду ограничений, связанных с механическими и динамическими свойствами языка. Модель протестирована на речевой базе, содержащей измерения поверхности языка для различных звуков. Ошибка аппроксимации контуров языка моделью во всех случаях оказалась в пределах погрешности измерений опытных данных.

Ключевые слова: теория речеобразования, математическая модель языка, уравнение упругого стержня, артикуляторный синтез речи, речевая обратная задача

DOI: 10.53921/18195822_2024_24_2_119

1. ВВЕДЕНИЕ

Математические модели языка необходимы для решения различных теоретических и практических проблем биологической акустики и речевых технологий: задачи артикуляторного синтеза речи, распознавания речи, речевой обратной задачи, а также различных биомедицинских акустических приложений [1–3]. В мировой литературе построено множество двухмерных и трехмерных моделей. Работы [4–6] посвящены трехмерным биомеханическим моделям, описывающим механику и кинематику мышц языка с помощью метода конечных элементов. Несмотря на большое теоретическое значение таких моделей, в настоящий момент они не вполне подходят для прикладных задач речевых технологий в силу крайней сложности управления такими моделями. Кроме того, задача описания механического взаимодействия между различными мышцами языка все еще является нерешенной проблемой для биомеханических моделей [7].

Гораздо более подходящим типом моделей языка для нужд речевых технологий являются так называемые функциональные модели. В их рамках контур языка в двух или трех измерениях аппроксимируется линейной комбинацией некоторых базисных контуров. Коэффициенты при контурах служат управляющими параметрами модели. В [8] контур языка в средне-сагиттальной плоскости описывается упругим стержнем в предположении малости его поперечных деформаций. Форма поверхности языка аппроксимируется линейной комбинацией нескольких собственных функций уравнения малых поперечных колебаний стержня (при некоторых заданных граничных условиях), а форма передней части языка – одной взвешенной собственной функцией того же уравнения. В [9,10] поверхность языка в двух и трех измерениях

аппроксимируется линейной комбинацией нескольких главных компонент рентгенографических измерений. Функциональные модели сравнительно просты в управлении и программной реализации; кроме того, они зачастую обеспечивают удовлетворительную точность аппроксимации измеренных контуров языка для различных звуков. Поэтому такие модели являются на текущий момент идеальными кандидатами для использования в практических задачах речевых технологий.

Наш многолетний опыт работы с различными функциональными моделями (в первую очередь, с моделью из [8]) показывает, что они, несмотря на несомненные достоинства, имеют ряд недостатков. Во-первых, они зачастую плохо подходят для аппроксимации артикуляций, при которых передняя часть языка поднимается вверх к твердому небу и значительно изгибается назад (например, при артикуляции английского звука R [11]). Во-вторых, эти модели, как правило, не накладывают никаких ограничений на длину или форму языка. В результате можно сгенерировать множество физиологически неправдоподобных или даже недопустимых контуров (например, контуры, имеющие излом в определенной точке на языке или же обладающие длиной, выходящей за допустимые физиологические границы). Наконец, в-третьих, функциональные модели, как правило, не накладывает никаких ограничений на траектории, описываемые точками на контуре языка во времени. В результате траектории могут допускать физиологически нереалистичные изломы в определенные моменты времени. Соответствующие конфигурации языка при этом будут демонстрировать скачкообразные изменения от одного временного кадра к другому.

Настоящая работа посвящена задаче построения функциональной модели языка, свободной от отмеченных недостатков известных моделей. Решение поставленной задачи достигается за счет 1) отказа от уравнения малых поперечных деформаций языка и перехода к нелинейному уравнению, допускающему сколь угодно большие поперечные деформации, 2) наложения ряда дополнительных статических и динамических ограничений на искомую форму языка для достижения физиологически адекватных язычных артикуляций.

2. ПОСТРОЕНИЕ МОДЕЛИ

Предположим, как и в [8], что конфигурация языка в средне-сагиттальной плоскости с достаточной степенью точности описывается гибким упругим стержнем длины l. Пусть EJ – жесткость языка, где E – модуль упругости языка, J – полярный момент инерции его поперечного сечения (для простоты мы предполагаем жесткость языка постоянной по пространственным координатам и по времени, хотя построенная ниже схема позволяет учесть и переменную жесткость). Введем параметр s^* – расстояние вдоль стержня, отсчитываемое от его начала (от корня языка), $0 \le s^* \le l$. Координата $s^* = 0$ соответствует корню языка, а $s^* = l$ – его кончику. Введем также нормированное расстояние вдоль стержня s, $s = s^*/\, l$, $0 \le s \le 1$.

Пусть исходная конфигурация языка описывается некоторой кривой C на плоскости (x, y). Введем функцию $\theta(s)$ – угол наклона касательной к кривой C, отсчитываемый от оси Ox против часовой стрелки. Под действием внешней распределенной силы с плотностью $q(s)$ язык будет деформирован и примет новое очертание, описываемое кривой C^*. Обозначим символом $v(s)$ угол наклона касательной к кривой C^*. Введем еще два обозначения: $\Delta(s) = v(s) - \theta(s)$, т.е. приращение угла наклона касательной к деформированному контуру языка C^* относительно угла наклона к исходной конфигурации языка C, и $\mu(s)$ – направление действия элементарной силы с модулем $q(s)ds$, отсчитываемое от оси Ox против часовой стрелки.

Уравнение, описывающее равновесие гибкого упругого стержня под действием внешней произвольной силы с плотностью $q(s)$, называется уравнением упругой линии. Это уравнение, применимое для определения конфигурации прямолинейных или криволинейных стержней при сколь угодно сильных изгибах, записывается следующим образом [12]:

$$\frac{d^2\Delta(s)}{ds^2} = -\frac{P_q(s)}{H}\sin\left[\theta(s) + \Delta(s) + \delta_q\left(\theta(s) + \Delta(s)\right)\right]. \quad (1)$$

Здесь введены обозначения:

$$\begin{aligned} P_q(s)\cos(\delta_q(s)) &= -\int_s^1 q(\xi)\cos(\mu(\xi))\,d\xi, \\ P_q(s)\sin(\delta_q(s)) &= \int_s^1 q(\xi)\sin(\mu(\xi))\,d\xi \quad . \end{aligned} \quad (2)$$

Из уравнения (2) следует, что функции $P_q(s)$ и $\delta(s)$ определяются так:

$$\begin{aligned} P_q(s) &= \sqrt{\left[\int_s^1 q(\xi)\cos(\mu(\xi))\,d\xi\right]^2 + \left[\int_s^1 q(\xi)\sin(\mu(\xi))\,d\xi\right]^2}, \\ \delta_q(s) &= \operatorname{arctg}\left(-\int_s^1 q(\xi)\sin(\mu(\xi))\,d\xi \Big/ \int_s^1 q(\xi)\cos(\mu(\xi))\,d\xi\right). \end{aligned}$$

В качестве граничного условия в точке $s = 0$ (корень языка) выберем жесткое прикрепление языка к подъязычной кости, т.е. $\Delta(\mathrm{s}) = 0$. В качестве граничного условия в точке $s = 1$ (кончик языка) выберем условие свободного конца, т.е. $\frac{d\Delta}{\mathrm{ds}}|_{s=1} = 0$. Нетрудно реализовать и шарнирное закрепление корня языка к подъязычной кости, однако, по результатам наших экспериментов, такое граничное условие не дает сколько-нибудь заметного роста точности аппроксимации измеренных конфигураций по сравнению с жестким закреплением.

Решая уравнение (1) с выбранными граничными условиями по известным данным $\theta(s)$, $q(s)$ и $\mu(s)$, мы получаем функцию $v(s) = \Delta(s) + \theta(s)$, по которой x-/y-координаты деформированной конфигурации C^* языка однозначно определяются из следующих соотношений:

$$\begin{aligned} \tfrac{x(s)}{l} &= \int_0^s \cos(v(\xi))\,d\xi, \\ \tfrac{y(s)}{l} &= \int_0^s \sin(v(\xi))\,d\xi\,. \end{aligned} \quad (3)$$

Аналитическое решение уравнения (1) возможно лишь в исключительных случаях для некоторых $\theta(s)$, $q(s)$ и $\mu(s)$ [13]. В большинстве случаев, представляющих интерес для практики, уравнение может быть решено только численно. К построению соответствующей численной схемы мы и переходим.

Для дальнейшего предположим, как и в работе [8], что исходная конфигурация языка длины l в начальный момент времени может быть с достаточной степенью точности аппроксимирована полуокружностью радиуса $R = l\ /\ \pi$. Введем две системы координат – $X\emptyset Y$ и $x\emptyset y$(Рис. 1). Обе системы имеют начало координат в точке, соответствующей корню языка, при этом система координат $x\emptyset y$ повернута относительно системы координат $X\emptyset Y$ на угол Θ_{tongue}. Также введем полярный угол φ, отсчитываемый от положительного направления оси Ox против часовой стрелки, $0 \le \varphi \le \pi$. Тогда координаты начальной полуокружности запишутся как $x = R + Rcos(\varphi)$, $y = Rsin(\varphi)$. Принимая во внимание, что для полуокружности $s^* = l$ - φR, $s = 1$ – $(\varphi R)/l$, получаем $x(s) = R$ - $Rcos(\pi s)$, $y(s) = Rsin(\pi s)$.

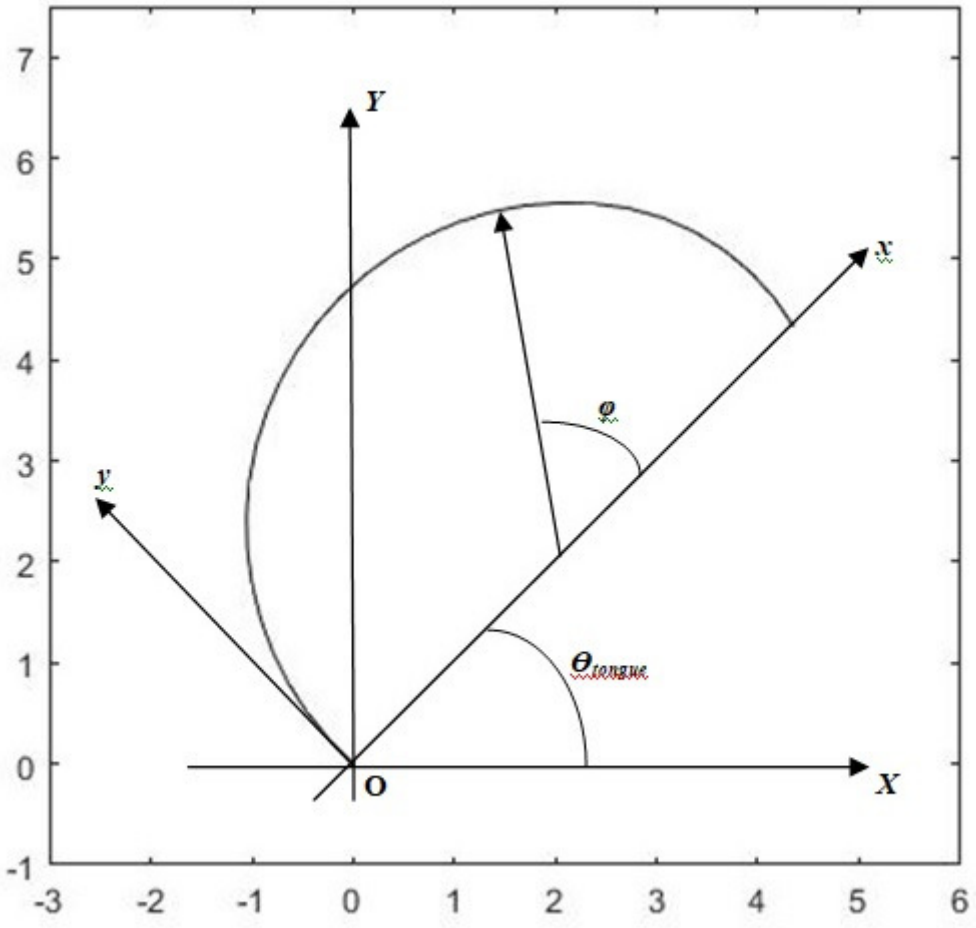

Рис. 1. Используемые системы координат.

Функция $\theta(s)$ определяется из следующей цепочки равенств: $\mathrm{tg}\theta(s) = \frac{\mathrm{dy}}{\mathrm{dx}} = \frac{y_s}{x_s} = \mathrm{ctg}(\pi\ s)$, откуда окончательно:

$$\theta(s) = \begin{array}{c} \frac{\pi}{2},\ s = 0, \\ \mathrm{arctg}(\mathrm{ctg}(\pi s)),\ 0 < s < 1, \\ -\frac{\pi}{2},\ s = 1. \end{array} \tag{4}$$

Разобьем интервал $0 \le s \le 1$ вдоль контура языка на последовательность равноотстающих отсчетов $s_1, s_2, \ldots, s_N$ с равномерным шагом h. Вводя обозначения $\Delta_i = \Delta(s_i)$, $q_i = q(s_i)$, $\mu_i = \mu(s_i)$, $\theta_i = \theta(s_i)$, $P_q(s_i) = (P_q)_i$, $\delta_q(s_i) = (\delta_q)_i$ аппроксимируя производные и интегралы в (1) и (2):

$$\frac{d^2\Delta(s)}{\mathrm{ds}^2} \approx \frac{\Delta_{i+1} - 2\Delta_i + \Delta_{i-1}}{h^2}, \tag{5}$$

и учитывая граничные условия $\Delta_1 = 0$, $\Delta_N = \Delta_{N\text{-}1}$, записываем (1) в виде нелинейной системы уравнений (6):

$$A\bar{\Delta} = \bar{f}. \tag{6}$$

Здесь A – трех-диагональная матрица следующей структуры:

$$A = \begin{pmatrix} -2 & 1 & 0 & 0 & \ldots & 0 \\ 1 & -2 & 1 & 0 & \ldots & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 \\ & & & \ddots & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots 1 & -2\ \ 1 \\ 0 & 0 & 0 & 0 & \ldots 0 & 1\ \ -1 \end{pmatrix}.$$

Соответствующие векторы определяются так (“T” – значок транспонирования):

$$\bar{\Delta} = (\Delta_2, ..., \Delta_{N\text{-}1})^T,$$

$$\bar{f} = (-\frac{(P_q)_2}{H}\sin[\theta_2 + \Delta_2 + \delta_q(\theta_2 + \Delta_2)], ..., -\frac{(P_q)_{N-1}}{H}\sin[\theta_{N-1} + \Delta_{N-1} + \delta_q(\theta_{N-1} + \Delta_{N-1})])^T.$$

Теоретически, численное решение системы (6) дает возможность определять контур языка по заданной внешней распределенной силе (прямая задача) либо по измеренному контуру языка вычислять соответствующую распределенную силу (обратная задача). Практически, работать с управляющим вектором внешних сил $\{q_\mathrm{i}, \mu_\mathrm{i}\}$, $i = 2, \ldots, N\text{-}1$ размерности $2\times(N\text{-}2) = 2N\text{-}4$ (число независимых управляющих параметров составляет от нескольких десятков до нескольких сотен) очень неудобно. Подбор параметров для решения прямой задачи, обычно осуществляемый вручную с помощью специального графического интерфейса [1], крайне

затруднителен в силу высокой размерности соответствующего пространства. Наложение статических и динамических ограничений на управляющий вектор для достижения физиологической адекватности контура языка теоретически возможно, однако практически приводит к системе нелинейных уравнений высокого порядка с нелинейными ограничениями в виде неравенств; решение такой задачи чрезвычайно трудоемко. Наконец, обратная задача определения вектора внешних сил по измеренному контуру языка из (6) оказывается некорректной. В частности, одной и той же форме языка может соответствовать множество различных векторов внешних сил. При этом ряд таких векторов представляет собой формальные решения, которые физиологически мало реалистичны или даже невозможны (например, в наших экспериментах мы иногда получали внешние силы с амплитудами в десятки кг; представляется нереалистичным, чтобы такие усилия создавались язычными мышцами). Таким образом, модель (6) требует конструктивной доработки.

Дальнейшее развитие модели опирается на следующую идею: пусть у нас имеется обширная база данных конфигураций языка, вычисленных с помощью (6) для некоторого заданного набора внешних сил. Применяя к этой базе процедуру кластеризации методом K-средних, получим множество из K классов и K соответствующих центроидов (каждый центроид соответствует одному определенному классу). В этом случае любой контур языка может быть аппроксимирован с необходимой точностью линейной комбинацией этих центроидов; коэффициенты при каждом центроиде будут новыми управляющими параметрами. Ожидается, что такой подход существенно облегчит нам решение как прямой, так и обратной задачи (в частности, значительно упростит наложение статических и динамических ограничений на управляющие параметры).

Пусть $\bar{x} = (x(s_1), ..., x(s_N), y(s_1), ..., y(s_N))^T = (x_1, ..., x_N, y_1, ..., y_N)^T$ - вектор размерности $2N$, описывающий поверхность языка в плоскости $x\mathrm{O}y$. Пусть далее $\left\{\bar{S}^1, ..., \bar{S}^K\right\}$ - набор из K центроидов; $\bar{S}^j = \left(X_1^j, ..., X_N^j, Y_1^j, ..., Y_N^j\right)^T, j = 1, ..., K$, X_i, Y_i – x-/y-координаты поверхности языка, соответственно (нижний индекс обозначает номер компоненты вектора, а верхний индекс – номер самого вектора). В рамках предлагаемой модели произвольный вектор $\bar{x}$ с необходимой точностью аппроксимируется линейной комбинацией центроидов $\left\{\bar{S}^1, ..., \bar{S}^K\right\}$, т.е.:

$$\bar{x} \approx \hat{x} = c_1\bar{S}^1 + c_2\bar{S}^2 + ... + c_K\bar{S}^K = \Phi\bar{c}. \tag{7}$$

Здесь Φ – матрица, столбцы которой соответствуют центроидам, $\bar{\mathrm{c}} = (\mathrm{c}_1, ..., c_K)^T$ - управляющий вектор. Уравнение (7) решает прямую задачу. Решение соответствующей обратной задачи для известной матрицы Φ может быть найдено как:

$$\tilde{c} = \underset{\bar{c}}{\operatorname{argmin}} \|\bar{x} - \Phi\bar{c}\|^2 = \left(\Phi^T\Phi\right)^{-1}\Phi^T\bar{x}. \tag{8}$$

(Здесь и далее используются следующие обозначения: для произвольных векторов $\bar{a} = (a_1, ..., a_N)^T, \bar{b} = (b_1, ..., b_N)^T$ их скалярное произведение есть $\langle\bar{a}, \bar{b}\rangle = a_1b_1 + ... + a_Nb_N$, $\|\bar{a}\|^2 = \langle\bar{a}, \bar{a}\rangle$- квадрат Евклидовой нормы).

Практически формула (8) неэффективна для решения обратной задачи без дополнительных ограничений на значения компонент управляющего вектора, поскольку вычисляемые с ее помощью управления зачастую порождают физиологически нереалистичные (и даже недопустимые) контуры языка. Кроме того, решение может оказаться неустойчивым относительно малых возмущений входных данных [14]. Для решения этих проблем вводятся специальные критерии (взвешенные стабилизирующие функционалы) и/или ограничения, необходимые для

того, чтобы модельная форма языка не только обеспечивала точную аппроксимацию измеренных данных, но и удовлетворяла требованиям физиологической адекватности. Коэффициент при каждом функционале определяет его вклад в суммарный критерий. Известно [15], что максимальное удлинение языка в экспериментах, когда испытуемый пытается достать кончиком языка свой подбородок, при этом помогая себе руками, не превышает 15% от исходной длины (в среднем по испытуемым, максимальное удлинение составляет 8.8%). В речи относительное удлинение заведомо меньше и составляет не более нескольких процентов. Поэтому первое физиологически обоснованное требование к модели заключается в том, чтобы длины порождаемых контуров языка мало отличались друг от друга. Пусть l – исходная длина языка, dl – некоторое допустимое изменение длины. Как правило, $dl << l$, так что можно положить $dl = 0$. Тогда соотношение (8) с учетом ограничения на длину языка можно записать следующим образом:

$$\begin{array}{l} \tilde{c} = \underset{\bar{c}}{\operatorname{argmin}} \|\bar{x} - \Phi\bar{c}\|^2 \\ \int\limits_0^1 \sqrt{\left(\frac{\mathrm{dx}(s)}{\mathrm{ds}}\right)^2 + \left(\frac{\mathrm{dy}(s)}{\mathrm{ds}}\right)^2}\,\mathrm{ds} = l. \end{array} , \tag{9}$$

Здесь $x(s)$, $y(s)$ - x-/y-координаты поверхности языка, определенные из (7).

Наличие квадратного корня под знаком интеграла в (9) существенно усложняет решение задачи. Этот недостаток можно преодолеть с помощью перехода от длины языка l к эквивалентной величине l_2: $l_2 = \int\limits_0^1 \left[\left(\frac{\mathrm{dx}(s)}{\mathrm{ds}}\right)^2 + \left(\frac{\mathrm{dy}(s)}{\mathrm{ds}}\right)^2\right]\mathrm{ds}$ (обращаем особое внимание на то, что величина l_2 не равна квадрату длины языка). В этом случае ограничение на длину языка запишется с помощью некоторой квадратичной формы, приведенной ниже.

Введем вспомогательные векторы:

$$\begin{array}{l} \bar{D}_x = (x_2 - x_1, ..., x_N - x_{N-1})^T , \\ \bar{D}_y = (y_2 - y_1, ..., y_N - y_{N-1})^T \end{array} . \tag{10}$$

Тогда:

$$\begin{array}{l} \int\limits_0^1 \left(\frac{\mathrm{dx}(s)}{\mathrm{ds}}\right)^2 \mathrm{ds} \approx \sum\limits_{i=1}^{N-1} [x_{i+1} - x_i]^2 = \langle \bar{D}_x, \bar{D}_x \rangle, \\ \int\limits_0^1 \left(\frac{\mathrm{dy}(s)}{\mathrm{ds}}\right)^2 \mathrm{ds} \approx \sum\limits_{i=1}^{N-1} [y_{i+1} - y_i]^2 = \langle \bar{D}_y, \bar{D}_y \rangle . \end{array} \tag{11}$$

Из (9) получаем:

$$\begin{array}{l} x_i = \sum\limits_{j=1}^{K} c_j X_i^j = \langle \bar{B}_X(i), \bar{c} \rangle, \bar{B}_X(i) = \left(X_i^1, ..., X_i^K\right)^T , \\ y_i = \sum\limits_{j=1}^{K} c_j Y_i^j = \langle \bar{B}_Y(i), \bar{c} \rangle, \bar{B}_Y(i) = \left(Y_i^1, ..., Y_i^K\right)^T \end{array} . \tag{12}$$

Отсюда

$$\bar{D}_x = \begin{pmatrix} \big(\bar{B}_X(2) - \bar{B}_X(1)\big)^T \\ \big(\bar{B}_X(3) - \bar{B}_X(2)\big)^T \\ \vdots \\ \big(\bar{B}_X(N) - \bar{B}_X(N-1)\big)^T \end{pmatrix} \bar{c} = B_X \bar{c}, \qquad \bar{D}_y = \begin{pmatrix} \big(\bar{B}_Y(2) - \bar{B}_Y(1)\big)^T \\ \big(\bar{B}_Y(3) - \bar{B}_Y(2)\big)^T \\ \vdots \\ \big(\bar{B}_Y(N) - \bar{B}_Y(N-1)\big)^T \end{pmatrix} \bar{c} = B_Y \bar{c}\,. \tag{13}$$

Подставляя (13) в (11) и вспоминая, что мы формулируем ограничение относительно l_2, а не l, окончательно получаем модель поверхности языка с ограничением на его длину:

$$\begin{aligned} &\tilde{c} = \underset{\bar{c}}{\operatorname{argmin}} \|\bar{x} - \Phi\bar{c}\|^2 \\ &\langle (B_X^T B_X + B_Y^T B_Y)\bar{c}, \bar{c}\rangle = \langle C_{\mathrm{XY}}\bar{c}, \bar{c}\rangle = l_2. \end{aligned}, \tag{14}$$

Использование эквивалентной длины языка l_2 вместо l очень эффективно: согласно результатам всех наших экспериментов, задача (14) решается примерно в 30 раз быстрее задачи (9) без какой-либо потери в точности аппроксимации язычной поверхности.

Теперь введем в рассмотрение статический критерий, требующий, чтобы контур языка был гладким. Уравнение (14) с учетом этого критерия перепишется так:

$$\begin{aligned} &\tilde{c} = \underset{\bar{c}}{\operatorname{argmin}} \left\{ \|\bar{x} - \Phi\bar{c}\|^2 + r\int\limits_0^1 \left[\left(\tfrac{d^2x(s)}{\mathrm{ds}^2}\right)^2 + \left(\tfrac{d^2y(s)}{\mathrm{ds}^2}\right)^2 \right] \mathrm{ds} \right\}, \\ &\langle C_{\mathrm{XY}}\bar{c}, \bar{c}\rangle = l_2 \end{aligned}. \tag{15}$$

Здесь r – параметр, определяющий вклад критерия гладкости языка в общее решение. Аппроксимируем вторые производные конечными разностями и вводим по аналогии с (13) вспомогательные матрицы:

$$B_X^2 = \begin{pmatrix} \big(\bar{B}_X(3) - 2\bar{B}_X(2) + \bar{B}_X(1)\big)^T \\ \big(\bar{B}_X(4) - 2\bar{B}_X(3) + \bar{B}_X(2)\big)^T \\ \vdots \\ \big(\bar{B}_X(N) - 2\bar{B}_X(N-1) + \bar{B}_X(N-2)\big)^T \end{pmatrix}, \qquad B_Y^2 = \begin{pmatrix} \big(\bar{B}_Y(3) - 2\bar{B}_Y(2) + \bar{B}_Y(1)\big)^T \\ \big(\bar{B}_Y(4) - 2\bar{B}_Y(3) + \bar{B}_Y(2)\big)^T \\ \vdots \\ \big(\bar{B}_Y(N) - 2\bar{B}_Y(N-1) + \bar{B}_Y(N-2)\big)^T \end{pmatrix}. \tag{16}$$

Полагая $C_{\mathrm{XY}}^2 = \big(B_X^2\big)^T B_X^2 + \big(B_Y^2\big)^T B_Y^2$, записываем (15) как:

$$\begin{aligned} &\tilde{c} = \underset{\bar{c}}{\operatorname{argmin}} \left\{ \|\bar{x} - \Phi\bar{c}\|^2 + r\langle C_{\mathrm{XY}}^2\bar{c}, \bar{c}\rangle \right\}, \\ &\langle C_{\mathrm{XY}}\bar{c}, \bar{c}\rangle = l_2 \end{aligned}. \tag{17}$$

Вводим множитель Лагранжа λ и строим целевую функцию:

$$\Omega^1(\bar{c},\lambda) = \|\Phi\bar{c} - \bar{x}\|^2 + r\langle C_{\mathrm{XY}}^2\bar{c},\bar{c}\rangle + \lambda(\langle C_{\mathrm{XY}}\bar{c},\bar{c}\rangle - l_2). \tag{18}$$

Дифференцируя это выражение по вектору c и приравнивая векторную производную нулю, получаем:

$$\tilde{c} = \left(2\Phi^T\Phi + \lambda\left(C_{\mathrm{XY}} + C_{\mathrm{XY}}^T\right) + r\left(C_{\mathrm{XY}}^2 + \left(C_{\mathrm{XY}}^2\right)^T\right)\right)^{-1} 2\Phi^T\bar{x}. \tag{19}$$

Коэффициент Лагранжа в (19) определяется из ограничения на эквивалентную длину языка l_2.

Рассмотрим теперь динамический критерий, требующий, чтобы траектории, описываемые точками на поверхности языка во времени, были достаточно гладкими. Для этого уравнение (17) нужно дополнить динамическим слагаемым вида $r_1\int\limits_0^T\|\frac{d^p\bar{x}}{\mathrm{dt}^p}\|^2\mathrm{dt}$. Здесь r_1 – вес динамического критерия, T – временной интервал, на протяжении которого действует этот критерий (обычно от нескольких сотен миллисекунд до нескольких секунд), t – время, p определяет конкретный тип динамического критерия. При $p = 1$ имеем – с точностью до мультипликативного постоянного вектора – критерий, минимизирующий кинетическую энергию точек языка [1]; случай $p = 2$ минимизирует ускорение точек на поверхности языка [16]; случай $p = 3$ соответствует условию минимальной резкости (minimal jerk) точек на язычной поверхности [17]. Чем выше p, тем более гладкими оказываются траектории точек на языке.

Далее ограничимся случаем $p = 1$. Пусть $\bar{x}^1, \bar{x}^2, ..., \bar{x}^M$- последовательность контуров языка, измеренных в дискретные моменты времени $n = 1, 2, \ldots, M$; $\bar{a}^1 = \Phi\bar{c}^1, \bar{a}^2 = \Phi\bar{c}^2, ..., \bar{a}^M = \Phi\bar{c}^M$- последовательность модельных конфигураций языка, подлежащих определению. Тогда задача формулируется следующим образом: необходимо найти последовательность контуров $\tilde{a}^1, \tilde{a}^2, ..., \tilde{a}^M$, удовлетворяющих следующему соотношению:

$$\left\{\tilde{a}^1, \tilde{a}^2, ..., \tilde{a}^M\right\} = \underset{\{\bar{a}^1,\bar{a}^2,...,\bar{a}^M\}}{\operatorname{argmin}} \left[\sum_{n=1}^{M}\|\bar{a}^n - \bar{x}^n\|^2 + r_1\sum_{n=2}^{M}\|\bar{a}^n - \bar{a}^{n-1}\|^2\right]. \tag{20}$$

Соотношение (20) означает следующее: рассматриваются всевозможные траектории, описываемые точками на поверхности языка $\{\ \bar{a}^1, \bar{a}^2, ..., \bar{a}^M\}$. Из этих траекторий выбирают такую, которая будет минимизировать критерий (20). В литературе по динамическим речевым обратным задачам встречается следующий подход к решению (20) (или схожих оптимизационных проблем) [1,18]: сначала для $n = 1$ определяют вектор из решения статической задачи, т.е. без учета динамического критерия. Затем последовательно для каждого значения $n = 2, 3, \ldots, M$ вычисляют векторы путем минимизации (20) (или аналогичного функционала) с уже известным из предыдущего шага вектором. Такой подход ошибочен: результирующие траектории не будут оптимальными относительно соответствующего критерия [19,20].

Алгоритм решения (20), вычисляющий контуры языка в дискретные моменты времени $n = 1, 2, \ldots, M$, и определяющий (в отличие от подходов из [1,18]) оптимальную траекторию задачи (20), построен в [21] методом динамического программирования. В силу громоздкости полученных соотношений этот алгоритм здесь не приводится. Соответствующие управляющие векторы для каждого язычного контура в каждый дискретный момент времени n находятся из (19).

Соотношения (1), (7), (19), (20) полностью определяют модель языка в средне-сагиттальном сечении.

Поскольку уравнение (1) определяет только статическую конфигурацию языка в средне-сагиттальной плоскости, но не динамику перехода от одной язычной формы к другой, особо остановимся на вопросе о том, как вычисляется управляющий вектор $\overline{c}(t)$ как функция времени t. Для решения обратной задачи восстановления поверхности языка по траекториям нескольких реперных точек функция $\overline{c}(t)$ определяется путем минимизации (20) (с учетом (18)) с помощью динамического программирования. Вместо динамического программирования могут быть использованы и другие методы из теории оптимального управления, например, алгоритмы на базе принципа максимума Понтрягина. Для решения прямой задачи порождения артикуляционных траекторий функция $\overline{c}(t)$ также определяется путем минимизации некоторого динамического критерия (с учетом ограничений на площади речевого тракта в контрольных сечениях и на координаты этих сечений, а также иных ограничений на форму языка). Пример решения задачи порождения динамики речевого тракта средствами динамического программирования содержится в [22]. В случае необходимости алгоритмы решения прямой и обратной задач могут быть дополнены различными моделями, связывающими $\overline{c}(t)$ и некоторые управляющие функции (например, может быть использована модель, в рамках которой $\overline{c}(t)$ есть отклик системы обыкновенных линейных дифференциальных уравнений второго порядка на возбуждающие кусочно-постоянные или кусочно-линейные команды). Такие модели будут учтены в качестве дополнительных ограничений на $\overline{c}(t)$. В любом случае, общая схема определения вектор-функции $\overline{c}(t)$ с помощью методов теории оптимального управления останется неизменной.

3. ТЕСТИРОВАНИЕ МОДЕЛИ И ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Тестирование модели происходило в два этапа. На первом создана база контуров языка. База подвергнута кластеризации методом K-средних, после чего определено необходимое число кластеров. На втором этапе соотношения (7), (19), (20) протестированы на реальных измерениях язычной поверхности для двух человек – носителей американского английского языка, произносивших различные звукосочетания. Рассмотрим каждый этап подробнее.

Исходная база контуров языка получена путем численного решения уравнения (6) с учетом соотношений (4), (5) и соответствующих граничных условий. Для этого сначала случайным образом было сгенерировано множество управляющих векторов $\{q_{\text{i}}, \mu_{\text{i}}\}$, $i = 1, \ldots, N$ (в наших экспериментах $N = 31$, так что размерность векторов $\bar{x}^k$ и всех центроидов $= 62$). При этом амплитуды внешних сил $\{q_{\text{i}}\}$ были ограничены диапазоном $[0, \ldots, 5$ г$]$, а углы $\{\mu_{\text{i}}\}$ определялись из схемы направления развиваемых усилий для мышц genioglossus superior, medialis, inferior, а также styloglossus, hyoglossus и longitudinalis superior, приведенной в [1]. Параметр Θ_{tongue} везде полагался равным 40^{o} и не варьировался. Результирующее множество внешних усилий составило 160 тыс векторов. Уравнение (6) с учетом соотношений (4), (5) и выбранных граничных условий решалось путем минимизации $\|A\overline{\Delta} - f\|^2 \to \min$ в среде MATLAB R2020a с помощью функции *fmincon*.

Полученная база контуров языка объемом 160 тыс векторов была подвергнута кластеризации методом K-средних. В качестве числа кластеров были выбраны следующие значения: $K = 4$, 8, 16, 32, 64, 128, 256, 512 и 1024 кластера. Для каждого K вычислялась ошибка аппроксимации $\sum\limits_{k=1}^{L} \|\bar{x}^k - \Phi(K)\,\bar{c}(K)\| / \sum\limits_{k=1}^{L} \|\bar{x}^k\|$ всей базы моделью (7); здесь L – общий объем базы (160 тыс векторов), $\bar{x}^k$- k-тый вектор из базы, $\Phi(K)$ – матрица, столбцы которой совпадают с K центроидами, $\bar{c}\,(K)$ - соответствующий управляющий вектор, вычисленный по формуле (8) (поскольку в данном случае сетки, на которых были заданы векторы $\bar{x}^k$ и цент-

роиды, совпадали, не было необходимости учитывать ограничения на длину языка и вводить дополнительный критерий гладкости контура). Для $K \geq 128$ ошибка аппроксимации составила сотые доли процента; при $16 \leq K \leq 64$ ошибка увеличилась с 0.09 ($K = 64$) до 0.87 $K = 8$ ошибка скачкообразно увеличилась до 62 $K = 4$ она составила более 98 модели (7) было выбрано значение $K = 16$. На Рис. 2 показаны некоторые из этих центроидов.

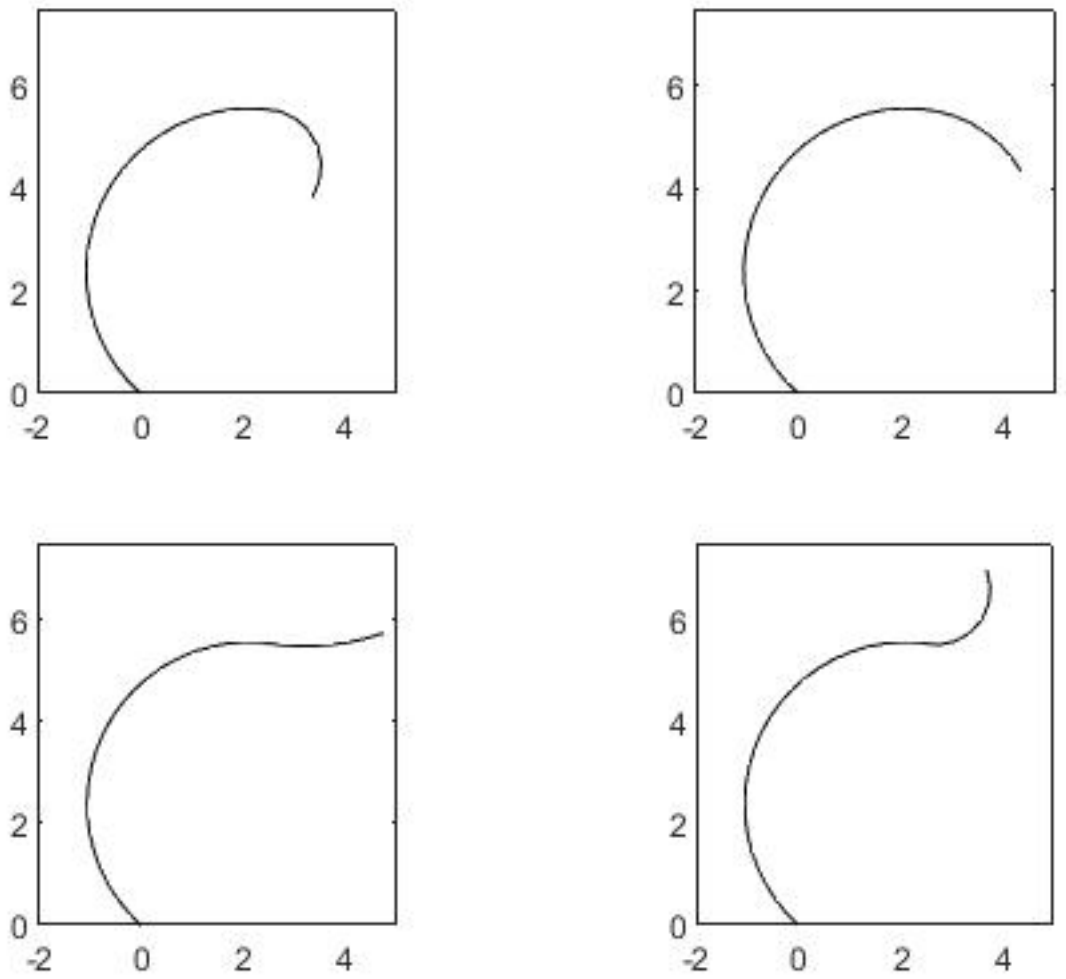

Рис. 2. Некоторые центроиды из базы контуров языка.

Тестирование точности модели осуществлялось на базе измерений, выполненных с помощью микролучевой рентгеноскопической установки в университете шт. Висконсин в 90-х годах [23]. В этой базе содержатся синхронные измерения траекторий четырех реперных точек на поверхности языка (а также на нижней челюсти и губах) и соответствующие акустические сигналы для нескольких десятков носителей американского английского языка, произносящих различные звуки, звукосочетания, слова и фразы по-английски. Первая реперная точка располагалась на расстоянии примерно 0.8 см от кончика языка, вторая – на расстоянии 2.5 см, третья – 4.4 см, наконец, четвертая – на расстоянии 6 см от кончика языка. Таким образом, четыре реперные точки позволяют оценить конфигурацию языка в его передней и средней

части при артикуляции различных звуков. Необходимо отметить, что приведенные значения являются средними по всем дикторам. Информация о точном расположении реперных точек для каждого диктора в базе отсутствует.

Размерность каждого измеренного вектора из микролучевой базы равна 8-ми (четыре x-координаты и четыре y-координаты для всех реперных точек). Поскольку, как указано выше, каждый центроид состоит из 31-ной точки (первая точка соответствует корню языка, 31-я – его кончику), было использовано следующее соответствие точек: первая реперная точка $\leftrightarrow$ 30-тая центроидная точка, вторая реперная точка $\leftrightarrow$ 27-я центроидная точка, третья реперная точка $\leftrightarrow$ 20-тая центроидная точка, четвертая реперная точка $\leftrightarrow$ 15-тая центроидная точка. Ограничение на длину языка, а также статический и динамический критерии при таком соответствии никак не меняются; уравнение (7) остается в силе, если в правой части брать не полноразмерные векторы центроидов, а усеченные векторы, компоненты которых равны компонентам центроидов в соответствующих точках.

Положение языка как твердого тела в каждый момент времени определяется как координатами его корня, так и управляющими параметрами нижней челюсти: ее углом раствора α_{jaw} и величиной горизонтального смещения x_{jaw}. Для необходимой калибровки сначала по измерениям координат двух реперных точек на нижнем резце и нижнем коренном зубе (относительно неподвижной системы координат, жестко связанной с верхней челюстью, с началом координат на основании верхних резцов) определялись α_{jaw} и x_{jaw}. Затем каждая измеренная реперная точка на поверхности языка с координатами $(x^j,\, y^j)^T$, $j = 1, \ldots, 4$, пересчитывалась в точку с координатами $\begin{pmatrix} x^j \\ y^j \end{pmatrix} = \begin{pmatrix} \cos\alpha_{\text{jaw}} & \sin\alpha_{\text{jaw}} \\ -\sin\alpha_{\text{jaw}} & \cos\alpha_{\text{jaw}} \end{pmatrix} \begin{pmatrix} x^j - x_{\text{jaw}} \\ y^j \end{pmatrix}$. Дальнейшие вычисления по формулам (7), (19), (20) осуществлялись с откалиброванными координатами.

Для тестирования были выбраны два диктора – диктор-мужчина (jw15 в соответствии с номенклатурой обозначений базы Westbury, 22 лет) и диктор-женщина (jw16, 20 лет). В качестве речевого материала использованы траектории четырех реперных точек на поверхности языка при произнесении дикторами звукосочетаний «гласный + гласный»: /IU/, /IA/, /UA/, /AU/, /AI/, /UI/ (Task15 в соответствии с номенклатурой обозначений базы Westbury), а также звукосочетаний «гласный + согласный + гласны» /ARA/, /AZA/, /ACHA/, /ASHA/, /AZHA/, /ASA/ (Task16). Выбор согласных обусловлен тем, что при их артикуляции кончик языка поднимается высоко вверх к альвеолам и даже может загибаться назад, создавая таким образом значительный изгиб в передней части (напомним, что, по нашему опыту, язычные конфигурации со значительным изгибом в передней части представляют для известных функциональных моделей большую проблему).

Для диктора-мужчины длина языка l полагалась равной 9.14 см (соответственно, $l_2 = 2.8$ см^2), для диктора-женщины $l = 8.82$ см ($l_2 = 2.6$ см^2). Коэффициенты r, r_1 при статическом критерии гладкости языка и динамическом критерии, соответственно, были экспериментально подобраны для звукосочетания /IU/ мужского и женского голоса; они составили $r = 0.05$, $r_1 = 0.01$. Эти значения в дальнейшем использовались для всех остальных звукосочетаний без изменения. Время действия динамического критерия полагалось равным общей длительности анализируемого звукосочетания (в среднем, 500 мсек). Частота дискретизации всех траекторий во времени составила 100 Гц. На всех графиках помимо контуров языка для большей ясности отображены также профили твердого неба, мягкого неба и нижней челюсти; соответствующие измерения хранятся в микролучевой базе.

На Рис. 3 в качестве примера показано несколько кадров модельной конфигурации языка для звукосочетания /ARA/ (верхняя строка соответствует диктору-мужчине, нижняя – диктору-женщине), а на Рис. 4 – несколько кадров для звукосочетания /IA/. Временной интервал между кадрами на этих графиках составляет 150 мсек. Интересно отметить, что дик-

торы используют различные артикуляторные тактики при произнесении /R/: в то время как диктор-мужчина поднимает кончик языка вверх к твердому небу, диктор-женщина, напротив, опускает кончик вниз и поднимает вверх все тело языка.

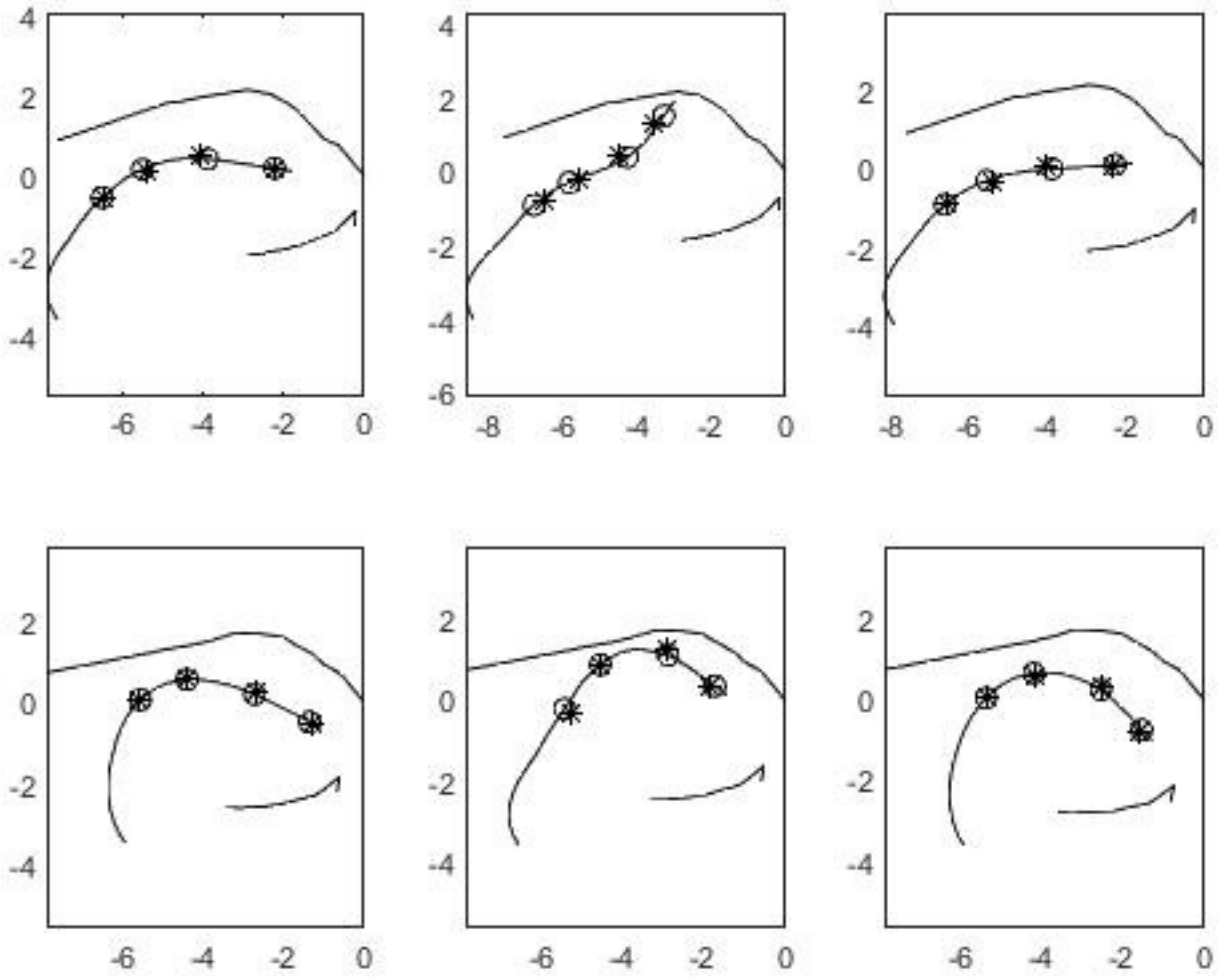

Рис. 3. Несколько кадров для звукосочетания /ARA/. Верхняя строка – диктор-мужчина, нижняя – диктор-женщина. "*" – измеренные реперные точки на поверхности языка, "o" – соответствующие точки, вычисленные по модели. Временной интервал между кадрами = 150 мсек.

Измеренные реперные точки отображены символом «*», соответствующие им точки на языке – символом «o». Погрешности аппроксимации измеренных реперных точек на поверхности языка нашей моделью для каждого звукосочетания и каждого диктора указаны в Табл. 1 (для каждого звукосочетания погрешность вычислялась как среднее (по всей последовательности временных кадров) значение величин $\sqrt{\frac{\sum_{k=1}^{4}\left(x^j_{\rm mes}(k)-x^j_{\rm calc}(k)\right)^2+\sum_{k=1}^{4}\left(y^j_{\rm mes}(k)-y^j_{\rm calc}(k)\right)^2}{\sum_{k=1}^{4}\left(x^j_{\rm mes}(k)\right)^2+\sum_{k=1}^{4}\left(y^j_{\rm mes}(k)\right)^2}}$ в (%). Здесь j – номер временного кадра, $(x,\ y)$ – x-/y-координаты измеренных реперных точек (подстроч-

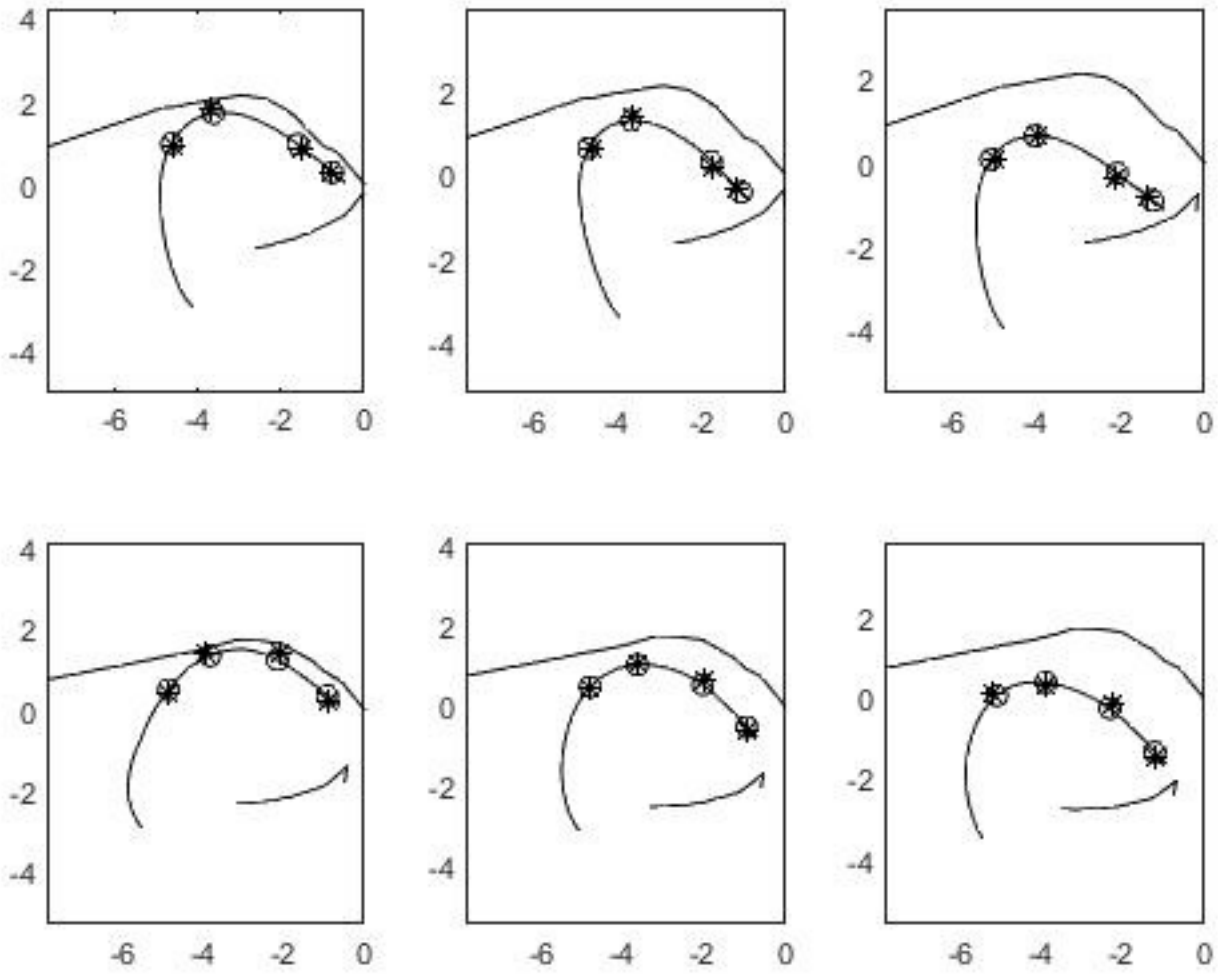

Рис. 4. Несколько кадров для звукосочетания /IA/. Верхняя строка – диктор-мужчина, нижняя – диктор-женщина. “*” – измеренные реперные точки на поверхности языка, “o” – соответствующие точки, вычисленные по модели. Временной интервал между кадрами = 150 мсек.

ное обозначение «mes» и соответствующих им точек, вычисленных по модели (подстрочное обозначение «calc».

Табл. 1. Погрешности аппроксимации поверхности языка моделью (%)

	jw15 (диктор-мужчина)	**jw16 (диктор-женщина)**
/IU/	2.9 %	4.3 %
/IA/	3.1 %	3.2 %
/UA/	2.8 %	3.6 %
/AU/	2.2 %	3.9 %
/AI/	3.1 %	2.9 %
/UI/	3.1 %	3.4 %
/ARA/	2.7 %	1.9 %
/ASA/	3.2 %	4.2 %
/AZA/	3.9 %	3.4 %
/ACHA/	2.3 %	3.7 %
/ASHA/	4.2 %	4.4 %
/AZHA/	4.2 %	3.7 %

При оценке результатов из Табл. 1 следует принимать во внимание погрешность измерения координат реперных точек, которая, согласно, [23], составляет порядка 3%. Дополнительную погрешность вносит отмеченное выше отсутствие информации о точном расположении реперных точек для каждого диктора (стандартное отклонение расположения точек на поверхности языка составляла порядка 1 см для точки на кончике языка и до 4.1 см для точки на середине языка). Поэтому установленное выше соответствие между измеренными реперными точками и центроидными точками, строго говоря, для разных дикторов должно быть различным. Поскольку в базе [23] отсутствовала информация о точном расположении реперных точек для каждого диктора (только усредненные значения по всем дикторам), мы для обоих дикторов использовали одинаковую схему соответствия точек. Дополнительные эксперименты показали, что вариация соответствия точек для каждого диктора приводит к дополнительному снижению ошибки аппроксимации (например, если для диктора-женщины в качестве соответствия второй реперной точке выбрать не 27-ю центроидную точку, а 25-тую, то ошибка аппроксимации для всех звукосочетаний снижается, при этом для /ASA/, /AZHA/, /ASHA/, /ACHA/ снижение составляет порядка 2%).

Таким образом, можно утверждать, что для обоих дикторов и всех исследованных звукосочетаний ошибка аппроксимации реперных точек на поверхности языка нашей моделью оказалась в пределах суммарной погрешности измерений.

4. ЗАКЛЮЧЕНИЕ

В работе построена математическая модель языка в средне-сагиттальном сечении. Модель представляет собой линейную комбинацию из шестнадцати центроидов, вычисленных по обширной базе решений уравнения упругой линии с заданными граничными условиями и распределенными внешними силами. Коэффициенты при центроидах являются управляющими параметрами модели. На коэффициенты наложено ограничение постоянства длины языка; кроме того, при вычислении коэффициентов используются критерии гладкости поверхности языка в пространстве (статический критерий) и во времени (динамический критерий). Ошибка аппроксимации реперных точек на язычной поверхности моделью для двух дикторов и различных звукосочетаний американского английского языка во всех случаях оказывается в пределах погрешности измерений.

Построенная модель позволяет с высокой точностью решать как прямую задачу (задачу порождения язычных артикуляций по заданным управляющим параметрам), так и обратную задачу (задачу восстановления контура языка по измеренным реперным точкам на язычной поверхности). Получающиеся при этом контуры во всех случаях являются физиологически правдоподобными.

СПИСОК ЛИТЕРАТУРЫ

1. В. Н. Сорокин, Речевые процессы. М.: ИД «Народное образование». 2012. 600 с.
2. A. Gómez, P. Gómez, D. Palacios, V. Rodellar, V. Nieto, A. Álvarez, and A. Tsanas, A Neuromotor to Acoustical Jaw-Tongue Projection Model With Application in Parkinson' Disease Hypokinetic Dysarthria // Front. Hum. Neurosci. 2021. Vol. 15. P. 1-20.
3. K. Al-hammuri, F. Gebali, I. Thirumarai Chelvan, A. Kanan, Tongue Contour Tracking and Segmentation in Lingual Ultrasound for Speech Recognition: A Review // Diagnostics. 2022, 12, 2811. P. 1-26.
4. V. Sanguineti, R. Laboissiere, Y. Payan Y. A control model of human tongue movements in speech // Biol Cybern. 1997. 77(1). P. 11–22.
5. S. Buchaillard, P. Perrier, Y. Payan. A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning // J. Acoust Soc Am. 2009. 126(4). P. 2033.
6. A. Gomez, F. Xing, D. Chan, D. Pham, J. Prince J. Motion estimation with finite-element biomechanical models and tracking constraints from tagged MRI. In: Wittek A, Joldes G, Nielsen PMF, Doyle BJ, Miller K, editors. Computational biomechanics for medicine. Cham: Springer Nature. 2017. P. 81–90.
7. A. Gomez, M. Stone, J. Woo, F. Xing, and J. Prince. Analysis of fiber strain in the human tongue during speech. In: Computer Methods in Biomechanics and Biomedical Engineering. Taylor & Francis Group. 2020. P. 1-11.
8. В. Н. Сорокин, Теория речеобразования. М.: Радио и связь. 1985. 313 с.
9. P. Badin, A. Serrurier, Three-dimensional Linear Modeling of Tongue: Articulatory Data and Models // Proceedings of the 7th ISSP. 2006. P. 395-402.
10. D. Beautemps, P. Badin, G. Bailly, Linear Degrees of Freedom in Speech Production: Analysis of Cineradio- and Labio-film Data and Articulatory-Acoustic Modeling // J. Acoust. Soc. Amer. 2001. 109(5). P. 2165-2180.
11. P. Ladefoged, I. Maddieson, The Sounds of the World's Languages. Blackwell Publishers. 1996.
12. Е. П. Попов, Теория и расчет гибких упругих стержней. М.: Наука. Гл. ред. физ.-мат.лит. 1986. 296 с.
13. Л. Д. Ландау, Е. М. Лифшиц, Теоретическая физика. Т. VII. Теория упругости. М.: Наука. Гл. ред. физ.-мат.лит. 1987. 248 с.
14. А. Н. Тихонов, А. С. Леонов, А. Г. Ягола, Нелинейные некорректные задачи. М.: Наука. 1995. 312 с.
15. W.-J. Kim, J.-B. Choi, J.-S. Park, S Lee, The Effects of Tongue Stretching Exercise on Tongue Length in Healthy Adults: a Preliminary Study // J. Phys. Ther. 2017. No. 29. P. 1929–1930.
16. T. Okadome, M. Honda, Generation of articulatory movements by using a kinematic triphone model // J. Acoust. Soc. Amer. 110 (1), July 2001. P. 453-463.
17. T. Flash, N. Hogan, The coordination of arm movements: an experimentally confirmed mathematical model // The Journal of Neuroscience, Vol. 5, No 7, July 1985. P. 1688-1703.
18. А.С. Леонов, И.С. Макаров, В.Н. Сорокин, А.И. Цыплихин, Артикуляторный ресинтез фрикативных // Информационные процессы. Том 4. No. 2. 2004. С. 141-159.

19. Р. Беллман, Р. Калаба, Динамическое программирование и современная теория управления. М.: Наука. Гл. ред. физ.-мат.лит. 1969. 120 с.

20. А. Брайсон, Хо Ю-Ши, Прикладная теория оптимального управления. М.: Издательство «Мир». 1972. 545 с.

21. И. С. Макаров, Об одном алгоритме оценки формантных частот на интервале сомкнутых голосовых складок // Речевые технологии. 2/2010. С. 45-65.

22. Kaburagi T., Honda M. A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes. Journal of the Acoustical Society of America, 1996, 99(5): 3154–3170.

23. J. Westbury, X-ray microbeam speech production database. User's handbook. Version 1. 1994.

On a mid-sagittal mathematical tongue contour model

I.S. Makarov

A mathematical tongue model in the mid-sagittal plane is constructed. A tongue contour is modeled as a flexible elastic beam, for which the elastic line equation for the given boundary conditions and the distributed input forces is numerically solved. The database of tongue contours collected from the solutions of the elastic line equation is clustered into 16 classes with the help of the K-means clustering algorithm. The resulting tongue contour is modeled as a linear combination of the clusters' centroids; coefficients for any centroid satisfy some constraints that correspond with mechanical and kinematic tongue properties. The model is tested on the database that contains tongue surface measurements for various sounds. The approximation error is within the data measurement discrepancy for all cases.

KEYWORDS: speech production theory, mathematical tongue model, elastic beam equation, articulatory speech synthesis, speech inverse problem.