Преобразование «цепочка фонем» → «речь» в динамических моделях: Обзор

Илья Сергеевич Макаров
ООО «БиометрикЛабс», Москва, Россия; im@biometriclabs.ru

Аннотация: Настоящий обзор посвящен динамическим моделям преобразования дискретной цепочки фонем в непрерывный речевой поток. Обсуждаются ключевые для современных динамических моделей понятия: артикуляционная модель, управляющие параметры, целевые артикуляции, артикуляционные жесты, принцип экономии произносительных усилий и проч. Излагаются результаты исследований специалистов Хаскинских лабораторий (артикуляционная фонология, task-dynamic-модель), а также японских исследователей (преимущественно Waseda University). Обзор иллюстрируется как модельными примерами, так и реальными артикуляционными измерениями на базе микролучевой рентгеноскопической установки.

Ключевые слова: автоматическая обработка речи, артикуляторная фонетика, компьютерная лингвистика, фонетика

Благодарности: Автор благодарен С. В. Князеву за очень ценные замечания, позволившие существенно улучшить статью.

Для цитирования: Макаров И. С. Преобразование «цепочка фонем» → «речь» в динамических моделях: Обзор. *Вопросы языкознания*, 2024, 1: 128–155.

DOI: 10.31857/0373-658X.2024.1.128-155

Phoneme sequence-to-speech conversion in dynamic phonological models: A survey

Ilya S. Makarov
BiometricLabs, LLC, Moscow, Russia; im@biometriclabs.ru

Abstract: This survey is devoted to dynamic models that model how a discrete phoneme sequence becomes converted to the corresponding continuous flow of articulations. The key concepts of modern dynamic models are discussed: articulatory model, articulatory parameters, goals and gestures, pronunciation effort economy principle, etc. The results of research conducted by specialists from Haskins Laboratories (articulatory phonology, task dynamic model), as well as by Japanese scientists (mostly from Waseda University) are presented. The survey is illustrated by both model examples and real articulatory X-ray microbeam measurements.

Keywords: articulatory phonetics, automatic speech processing, computational linguistics, phonetics

Acknowledgements: The author is grateful to S. V. Knyazev for very valuable comments that helped us improve the article significantly.

For citation: Makarov I. S. Phoneme sequence-to-speech conversion in dynamic phonological models: A survey. *Voprosy Jazykoznanija*, 2024, 1: 128–155.

DOI: 10.31857/0373-658X.2024.1.128-155

Введение

Основная задача динамических фонологических моделей[1] (например, моделей, разработанных в рамках **артикуляционной фонологии**) заключается в построении алгоритмов перехода от **дискретной цепочки фонем** (а также некоторых глубинных просодических и временных представлений текста, соответствующего данной цепочке фонем) к **непрерывному речевому потоку**. При этом под **речевым потоком** может пониматься как непрерывная последовательность конфигураций речевого тракта (т. е. траектории, описываемые точками на поверхностях языка, нижней челюсти, губ и других органах речевого аппарата при совершении различных артикуляционных движений), так и соответствующий этим артикуляциям речевой сигнал. С теоретической точки зрения последовательность конфигураций речевого тракта (а также некоторая дополнительная информация, связанная с акустическими источниками звука в речевом тракте) однозначно определяет соответствующий акустический сигнал [Леонов и др. 2003; 2004; 2005; Макаров 2009; 2011; Sorokin et al. 2005]. Поэтому в дальнейшем под речевым потоком будет пониматься только последовательность конфигураций речевого тракта.

Отметим некоторые основные проблемы, с которыми (в теории) динамические фонологические модели обязаны успешно справляться[2]: 1) порождение всех коартикулляционных явлений[3], наблюдаемых в реальной речи, 2) моделирование речевого потока как для **полного** (по Щербе [1974]), так и для **неполного** типов произношения, 3) генерация потока для произвольного анатомического строения речевого аппарата (в том числе, для таких случаев, как отсутствие зубов или наличие явных речевых патологий типа заячьей губы), 4) учет положения человека в пространстве (говорит ли человек стоя, сидя, лежа, с запрокинутой головой, при медленной или быстрой ходьбе, на бегу и т. д.), 5) порождение всех компенсационных артикуляционных явлений, возникающих из-за дополнительных внешних ограничений для тех или иных органов речевого аппарата (например, человек говорит с сигаретой в зубах, с конфетой во рту, под местной анестезией в стоматологическом кабинете и т. д.).

При построении динамических фонологических моделей исключительное теоретическое и практическое значение имеют базы, содержащие в себе измерения движений артикуляционных органов разных людей в различных экстралингвистических условиях при произнесении различного речевого материала. Измерения могут быть выполнены с помощью различных методов, например кинорентгеносъемки, электро-магнитной артикулографии, компьютерной томографии речевого тракта и др. [Кодзасов, Кривнова 2001: 93–97]. В настоящем обзоре в качестве базы измерений используются артикуляционные записи речевого тракта, полученные с помощью микролучевой рентгеноскопической установки в университете штата Висконсин [Westbury 1994]. В этой базе записей содержатся синхронные измерения траекторий нескольких миниатюрных датчиков (называемых в дальнейшем **реперными точками**)[4], приклеенных к различным участкам поверхности языка,

---

[1] Подробнее об истории, принципах и подходах динамической фонологии см. [Кодзасов, Кривнова 2001: гл. 10].

[2] Дальнейший список проблем, очевидно, неполон.

[3] Под коартикуляцией здесь и везде далее понимается влияние артикуляций соседних звуков друг на друга, планируемое на уровне моторной программы высказывания. Коартикуляция (как и все фонетические явления, описываемые динамическими фонологическими моделями) осуществляется **после** завершения действия всех фонологических (лингвистических) правил — ассимиляции, редукции гласных, оглушения / озвончения конечных согласных и т. д. О разграничении фонологических правил и синхронных фонетических изменений, а также об их иерархии см., например, [Князев 1999; 2001; Кодзасов, Кривнова 2001: 67 и сл.; Князев 2004].

[4] Вообще говоря, под реперной точкой можно понимать произвольную точку наблюдения / измерения на поверхности того или иного артикуляционного органа.

нижней челюсти и губ, а также соответствующие акустические сигналы для нескольких десятков носителей американского английского языка, произносящих различные звуки, звукосочетания, слова и фразы по-английски. На рис. 1 показано расположение и нумерация этих реперных точек: 1 — точка на верхней губе, 2 — точка на нижней губе, 3 — точка на нижнем переднем резце, 4 — точка на нижнем жевательном зубе, 5–8 — четыре точки на поверхности языка. Для удобства восприятия реперные точки отображены на фоне некоторых модельных конфигураций языка и губ. Особо отметим, что в базе хранится информация только о восьми точках (а также об индивидуальных анатомических размерах каждого диктора: форма твердого неба, размеры верхней и нижней челюсти, а также зубов и губ, размеры черепа, координаты точки прикрепления нижней челюсти к черепу и т. д.), а не о всей поверхности языка и губ[5]. На рис. 2 и 3 в качестве примера показаны измеренные траектории точки 2 (нижняя губа), точки 3 (нижний передний зуб) и точек 5-8 (поверхность языка) при произнесении звукосочетания /IA/ носителем американского варианта английского языка. Измеренные траектории отрисованы линией «x-». Для удобства восприятия на траектории наложены модельные конфигурации речевого тракта. На рис. 2 модельная речевая конфигурация соответствует фазе выдержки начального гласного /I/; на рис. 3 конфигурация соответствует фазе выдержки конечного гласного /A/.

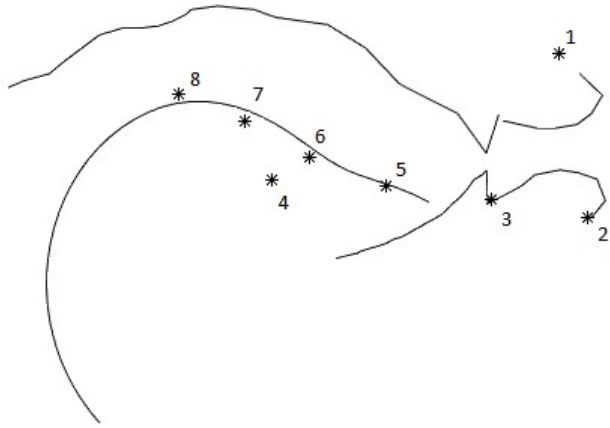

Рис. 1. Расположение реперных точек из базы данных [Westbury 1994] на фоне некоторой модельной конфигурации языка и губ

[5] Вопрос о том, каким образом по информации о расположении нескольких реперных точек восстановить всю поверхность языка и губ, является краеугольным при анализе измерений, выполненных с помощью микролучевого рентгеноскопа или электромагнитного артикулографа. Этот вопрос представляет значительные технические сложности и здесь не рассматривается. Некоторые подходы к решению содержатся в [Kaburagi, Honda 1994; Сорокин 2012: 378 и сл.; Wang et al. 2014; Макаров 2023].

Ситуация усугубляется тем, что реперные точки прикрепляются к поверхностям речевого тракта только в его передне-средней части. Информация о нижней части тракта (задняя часть языка, средняя и нижняя часть глотки, надгортанник и эпиларинкс) отсутствует. Можно ли только по информации о траекториях нескольких реперных точек, а также по синхронно записанному акустическому сигналу восстановить всю конфигурацию речевого тракта — от голосовой щели до губ? Этот вопрос исключительно сложен и еще далек от своего полного решения. Один из подходов к решению этой проблемы предложен в [Макаров 2005].

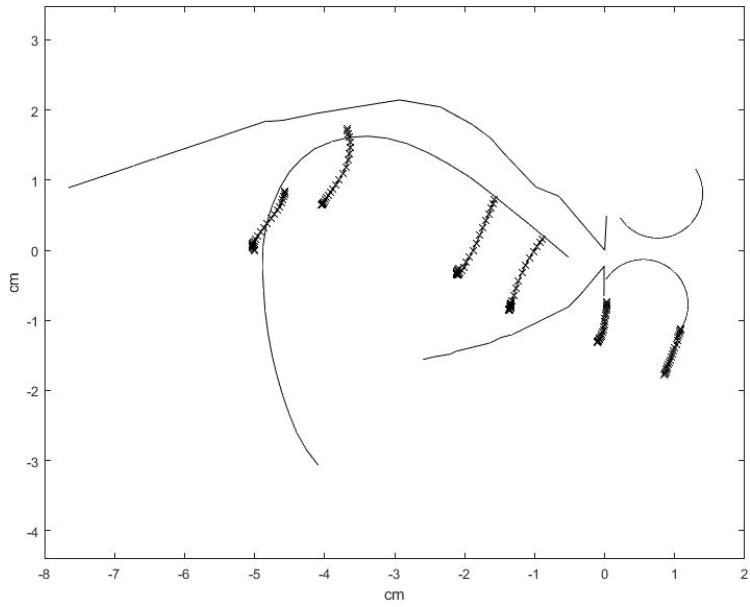

Рис. 2. Измеренные траектории реперных точек при артикуляции звукосочетания /IA/ на фоне модельной конфигурации речевого тракта, соответствующей фазе выдержки начального гласного /I/

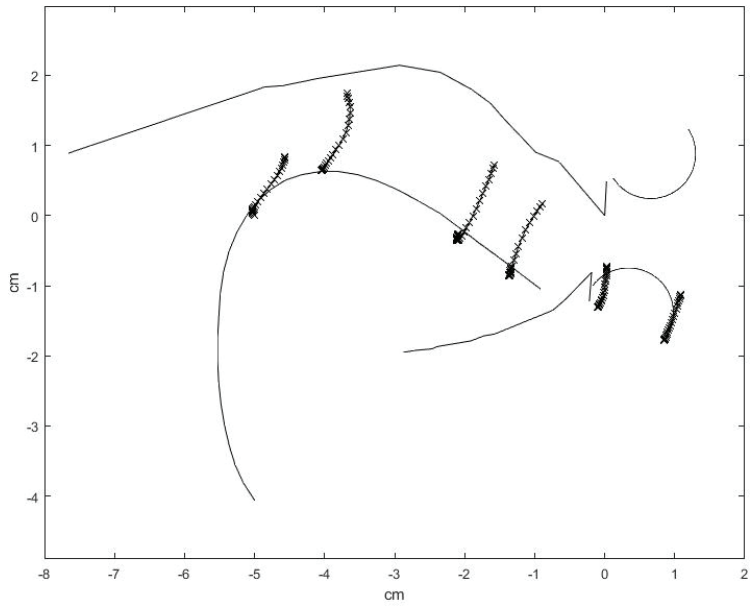

Рис. 3. Те же траектории, что и на рис. 2, с наложенной модельной конфигурацией речевого тракта, соответствующей фазе выдержки конечного гласного /A/

При написании обзора автор столкнулся с тремя трудностями. Первая трудность заключается в том, что разные динамические модели нередко используют различную

терминологию для описания одних и тех же (или очень близких) понятий и явлений. Наблюдается и обратная тенденция — иногда один и тот же термин в разных работах используется для обозначения весьма различных понятий. Поэтому автор предпринял попытку унифицировать терминологический язык для более компактного описания моделей, а также для облегчения задачи сопоставления различных моделей друг с другом. В табл. 1 приведена сводка основных терминов, используемых различными динамическими моделями, а также предлагаемый автором унифицирующий термин (объяснения всех терминов приводятся ниже; насколько при этом термин является удачным — предлагается судить читателю). Вторая трудность видится в том, что динамические модели базируются на обширном массиве знаний о физиологических, механических и кинематических свойствах речевого аппарата и используют для описания этих знаний весьма сложный математический аппарат (в первую очередь теорию оптимального управления, а также аппарат, составляющий математический фундамент теории упругости и сопротивления материалов). Поскольку нереалистично ожидать от потенциального читателя настоящего обзора знакомства с данными математическими дисциплинами, автор во всех случаях избегал каких-либо математических формул и старался объяснять необходимые понятия неформально. Наконец, третья трудность заключается в том, что в русскоязычной фонетической литературе для ряда терминов (например, «целевая артикуляция», «артикуляционный жест») закрепилось понимание, которое не вполне соответствует (а иногда — прямо противоречит) тому пониманию, которое вкладывали в эти термины классики динамических фонологических моделей (создатели артикуляционной фонологии и пр.). В настоящем обзоре автор приводит только классические толкования терминов и никак не учитывает их возможные русскоязычные реинтерпретации.

Таблица

Терминологическая сводка

Предлагаемый русскоязычный термин	Англоязычные аналоги
Динамическая фонологическая модель	Dynamic articulatory model, Dynamic phonological model, Task-dynamics model[6], Trajectory formation model
Управляющий (артикуляционный) параметр	State variable, Articulatory parameter, Articulatory variable, Tract variable[7] Model articulator variable, Articulatory degree-of-freedom Articulatory coordinate Vocal-tract coordinate

[6] В литературе термин «task dynamics model» обычно используется для обозначения динамической фонологической модели, разработанной специалистами Хаскинских лабораторий (Haskins Labs). **Task dynamics model** является ключевым модулем в составе **артикуляционной фонологии** (articulatory phonology), отвечающим за вычисления артикуляционных траекторий по заданной дискретной цепочке фонем.

[7] В ряде работ термин «tract variable» используется как в значении «управляющий параметр», так и в значении «реперная точка». Во многих работах представителей Хаскинских лабораторий (в том числе, в классической статье [Saltzman, Munhall 1989]) под «tract variable» понимают место и степень сужения в речевом тракте, а для обозначения управляющего артикуляторного параметра используют термин «model articulator variable». При чтении зарубежных работ, посвященных динамическим фонологическим моделям, надо всегда иметь в виду такую полисемию

Предлагаемый русскоязычный термин	Англоязычные аналоги
Реперная точка	Tract variable
Артикуляционная модель	Vocal-tract model, Articulatory model
Целевая артикуляция (= артикуляционная цель)	(Phoneme-specific) motor task, Phonemic articulatory target, Phoneme-specific articulatory goal, Phoneme-specific invariant feature(s)
(Артикуляционный) признак	Tract variable[8] Tube feature
Артикуляционный жест	Articulatory gesture
Партитура	Gestural score
Интервал активации	Activation interval, Interval of active gestural control
Артикуляционные ограничения	Articulatory constraints
Критерий оптимальности	Optimality criterion, Cost function

В настоящем обзоре используются два типа примеров, иллюстрирующих те или иные аспекты динамических алгоритмов: модельные примеры, полученные с помощью некоторых алгоритмов расчета конфигураций речевого тракта и соответствующей акустики, и примеры реальных артикуляционных траекторий. Все модельные примеры выполнены автором; используемые артикуляционные и акустические модели с разной степенью подробности описаны в [Макаров, Сорокин 2004; Баден и др. 2005; Макаров 2009; 2011]. Артикуляционные траектории взяты автором из базы микролучевых рентгеноскопических измерений [Westbury 1994].

1. Основной понятийный аппарат динамических моделей

Рассмотрим основные понятия, присущие всем динамическим фонологическим моделям.

Под **фонемой** здесь и везде далее понимается символьная сегментная единица для записи означающих Лексикона[9] данного языка после действия всех фонологических правил. В традиционной фонологии нет прямого аналога этому термину. Наверное, следовало бы придумать какой-либо особый термин, например «пост-фонема» или «фонема поверхностного уровня», однако ввиду чрезвычайно широкой распространенности термина «фонема» именно в таком понимании в динамических моделях мы здесь и везде далее будем употреблять только этот термин. Мы предполагаем, что набор фонем, необходимый и достаточный для описания Лексикона данного языка, уже определен тем или иным

и каждый раз по контексту догадываться, в каком именно значении используется тот или иной термин.

[8] В таком значении термин «tract variable» употребляется только в работах специалистов Хаскинских лабораторий (артикуляционная фонология, task dynamics model).

[9] Под Лексиконом здесь и далее понимается Словарь всех словоформ данного языка (такое понимание типично для многих работ в области автоматического распознавания речи [Huang et al. 2001])

способом[10]. Особо подчеркнем, что установление фонетических оппозиций и чередований в данном языке, определение количества фонем и отнесение того или иного звука к определенной фонеме — все эти задачи не имеют никакого отношения к проблематике динамических моделей; динамические модели предполагают, что эти проблемы уже решены на более глубоком фонологическом уровне.

Фонема — не единственная символьная сегментная единица, которая может быть использована в динамических моделях. Другими кандидатами для записи означающих Лексикона могут быть **дифон** (сегментная единица, стоящая после другой сегментной единицы) или же чрезвычайно популярный в системах автоматического распознавания речи **трифон** (сегментная единица, стоящая между двумя другими сегментными единицами) [Kaburagi, Honda 2001; Okadome, Honda 2001]. Несмотря на существенную разницу в числе фонем, дифонов и трифонов для данного языка (несколько десятков фонем, сотни дифонов и тысячи трифонов), динамические модели на базе этих единиц дают близкие результаты (хотя, разумеется, техническая реализация фонемных, дифонных и трифонных моделей может существенно различаться).

Другие возможные сегментные единицы (например, слог) мы не рассматриваем.

Под **динамической фонологической моделью** понимается алгоритм, который по заданной дискретной цепочке фонем данного языка (а также по дополнительной временной разметке текста, соответствующего данной цепочке фонем) вычисляет непрерывную последовательность конфигураций речевого тракта в двух или трех измерениях[11], соответствующую этой цепочке. Под временной разметкой текста понимается указание моментов времени, в которых должны быть достигнуты **целевые артикуляции** для каждой фонемы (определение «целевой артикуляции» см. ниже).

Динамические модели разделяются на **детерминированные** и **статистические**. В рамках первого класса моделей непрерывная последовательность конфигураций речевого тракта определяется с помощью различных уравнений теории колебаний или теории упругости, использующих в качестве параметров механические и кинематические свойства различных артикуляционных органов. Второй класс моделей строится на основе некоторой обширной базы артикуляторных измерений, сегментированной на составляющие фонемы. С помощью этой базы определяются статистические (вероятностные) правила перехода от заданной цепочки фонем к соответствующим артикуляциям. Для произвольной входной цепочки фонем соответствующий речевой поток вычисляется на базе этих правил. Настоящий обзор посвящен главным образом детерминированным моделям, которые и более популярны, чем статистические, и имеют большую объяснительную силу.

Управляющими артикуляционными параметрами (или просто управляющими параметрами) называют набор постулируемых координат речевого тракта, изменение которых вызывает соответствующее изменение всей артикуляционной конфигурации. Примером управляющего параметра служит угол поворота нижней челюсти относительно ее точки прикрепления к черепу: увеличивая значение этого угла, мы изменяем текущее положение нижней челюсти, языка и нижней губы (= чем больше угол поворота, тем шире

[10] Обычно состав фонем для динамических моделей устанавливается исходя из системы фонетических оппозиций в данном языке. При этом сложные для традиционной фонологии случаи, как правило, выделяются в отдельные фонемы. Например, в рамках динамической модели для русского языка в состав гласных фонем в таком понимании войдут, наряду с /И, У, Э, О, А/, еще /Ы/ и /ə/ (шва). Особо отметим, что выделение /Ы/ и /ə/ как особых фонем (в таком понимании!) диктуется не функциональными соображениями (т. е. анализом фонетических оппозиций и чередований), но соображениями удобства при решении задачи порождения артикуляций по дискретной цепочке символов.

[11] Далее все примеры приводятся для конфигурации речевого тракта только в двух измерениях, т. е. в средне-сагиттальной плоскости (это соответствует привычным для фонетистов картинкам состояния речевого аппарата).

открывается рот). Другим примером управляющих параметров служат горизонтальная и вертикальная координаты корня языка. Изменяя текущие значения этих параметров, мы меняем положение тела языка во рту (для гласных звуков горизонтальная координата корня языка изменяет **ряд** гласного, а вертикальная координата — **подъем**). Еще примеры управляющих параметров: угол поворота небной занавески относительно точки прикрепления к твердому небу (чем больше угол, тем сильнее носовая полость акустически подключена к речевому тракту), вертикальное расстояние между губами (когда расстояние равно нулю, на губах образуется смычка), горизонтальное смещение губ (чем больше смещение, тем сильнее лабиализация), вертикальная и горизонтальная координаты кончика языка и т. д. Выбор тех или иных координат определяется опытом и вкусами исследователя, а также решаемыми задачами. В некоторых современных биомеханических моделях речевого тракта управляющими параметрами служат механические и кинематические параметры мышц, определяющих текущее положение артикуляционных органов; см., например, [Gomez et al. 2020]. Несмотря на огромную теоретическую значимость таких моделей, ими очень сложно управлять; кроме того, с такими моделями связан ряд принципиальных и до сих пор не решенных физиологических и математических сложностей. Поэтому в подавляющем большинстве современных динамических моделей используются гораздо более простые наборы управляющих параметров типа тех, которые были упомянуты выше.

Фундаментальным понятием для всех динамических фонологических моделей является **артикуляционная модель**. Под артикуляционной моделью понимается набор управляющих параметров и алгоритм, который для каждого значения каждого управляющего параметра вычисляет соответствующую конфигурацию речевого тракта. На рис. 4 показаны две конфигурации речевого тракта, вычисленные для двух различных значений угла поворота нижней челюсти. На этом рисунке управляющие параметры показаны в виде ползунков; меняя положение того или иного ползунка, мы меняем значение соответствующего управляющего параметра. Специальный алгоритм по текущим положениям ползунков вычисляет конфигурацию речевого тракта, которая затем отображается на экране (вообще неформально всякую артикуляторную модель можно представлять себе как графический интерфейс типа, показанного на рис. 4 (с. 136), содержащий набор ползунков-управлений и алгоритм для отрисовки всей конфигурации речевого тракта после изменения положения произвольного ползунка).

Существует большое количество артикуляционных моделей, отличающихся друг от друга (иногда очень существенно) используемыми управляющими параметрами и алгоритмами расчета соответствующей конфигурации речевого тракта. Читателю, интересующемуся тем, как именно строятся такие модели, можно рекомендовать подробное описание этого процесса, изложенное в [Сорокин 1992: гл. 8; 2012: 96–105]. Мы не будем углубляться в этот вопрос; отметим только, что в разных динамических моделях могут быть использованы (и фактически используются) разные артикуляционные модели.

Дальнейшее изложение в значительной мере опирается на следующее утверждение: для произвольного языка задание 1) сужений в некоторых областях речевого тракта (а также площади просвета между голосовыми складками), 2) общей длины тракта от губ до гортани, 3) задания временного контура подсвязочного давления необходимо и достаточно для генерации основных аэродинамических и акустических явлений, выполняющих в данном языке смыслоразличительную функцию. Утверждение весьма хорошо согласуется с экспериментальными данными и результатами моделирования аэродинамических [Сорокин 1992: 134–143; Hanson, Stevens 2002] и акустических [Fant 2001; Story, Bunton 2011; 2019] процессов речевого тракта в частотном диапазоне до (примерно) 4 000 Гц.

Очень важным понятием в динамических фонологических моделях является понятие **целевой артикуляции** (или **артикуляционной цели**) для каждой фонемы. Под целевой артикуляцией фонемы понимаются значения площадей поперечных сечений[12] речевого

[12] С акустической точки зрения речевой тракт представляет собой трубу переменного поперечного сечения и переменной длины. Различным конфигурациям тракта при этом будут соответствовать

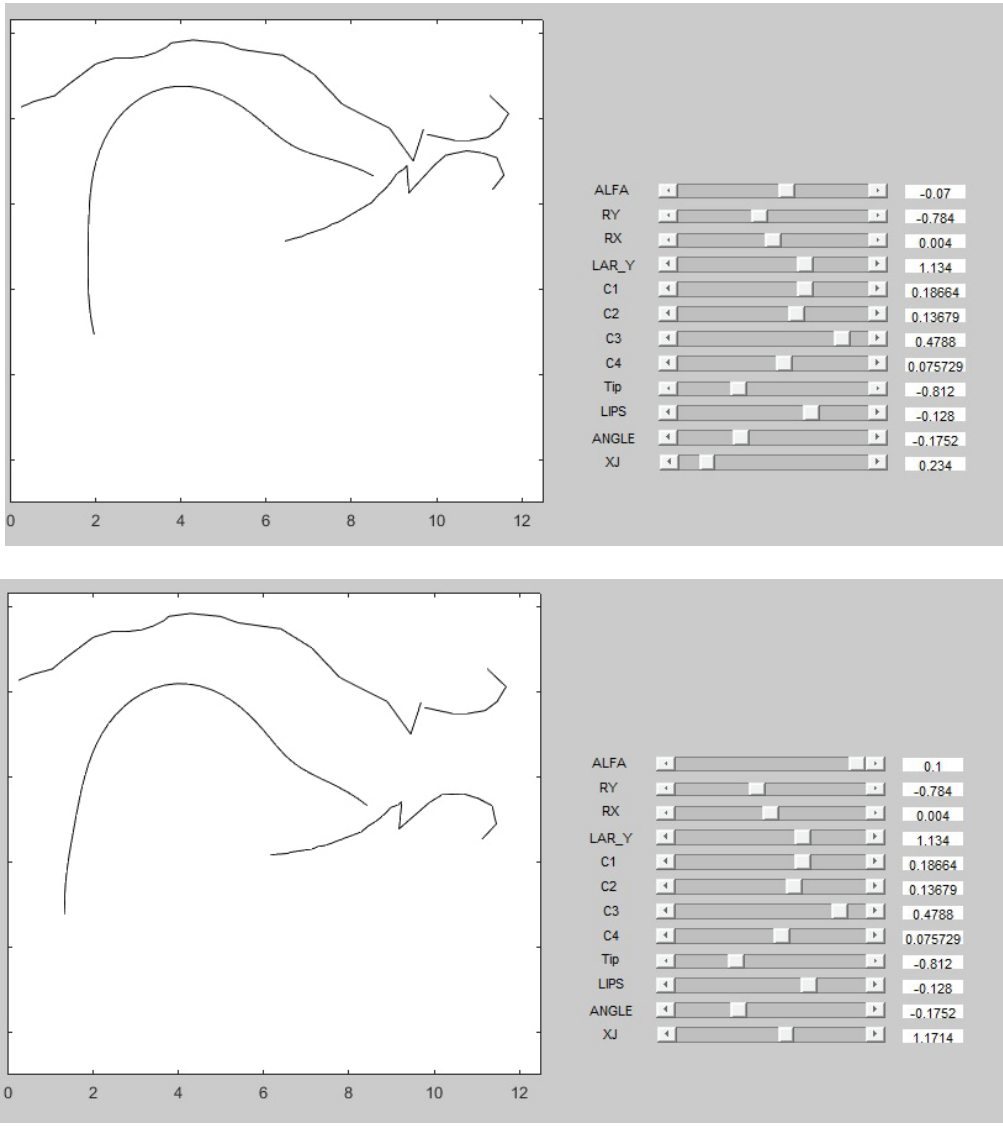

Рис. 4. Две конфигурации речевого тракта, построенные с помощью некоторой артикуляционной модели. Конфигурации отличаются друг от друга значениями угла поворота нижней челюсти (= степенью раскрытия рта) (условное название соответствующего ползунка = ALFA)

тракта в определенных контрольных областях речевого тракта, а также приращение длины речевого тракта относительно некоторого нейтрального состояния артикуляционного аппарата (под нейтральным состоянием обычно понимают состояние покоя речевого тракта: человек молчит, челюсти и губы сомкнуты, никаких движений, в том числе жевательных,

различные распределения площадей поперечных сечений этой трубы от гортани до губ (подробнее см. [Кодзасов, Кривнова 2001: 115–118]).

во рту не происходит)[13]. Иначе говоря, для каждой фонемы ее целевая артикуляция определяется заданием а) координат контрольных областей (координата отсчитывается вдоль средней линии тракта от гортани к губам), б) значений площадей поперечных сечений тракта в этих координатах, в) изменения общей длины тракта от гортани до губ относительно длины в нейтральном состоянии. Для дальнейшего обсуждения условно выделим в речевом тракте набор из восьми контрольных областей[14] (рис. 5): 1-я область — область между губами (губная область), 2-я область — область между нижней губой и верхними передними зубами (губно-зубная область), 3-я область — область между кончиком языка и верхними передними зубами (переднеязычно-зубная область), 4-я область — область между передней частью языка и альвеолами (переднеязычно-альвеолярная область), 5-я область — область между передне-средней частью языка и средней частью твердого неба (палатальная область), 6-я область — область между серединой языка и мягким небом (велярная область), 7-я область — область между задней частью языка и задней стенкой речевого тракта (заднеязычная область), 8-я область — область между небной занавеской и задней стенкой речевого тракта (назальная область).

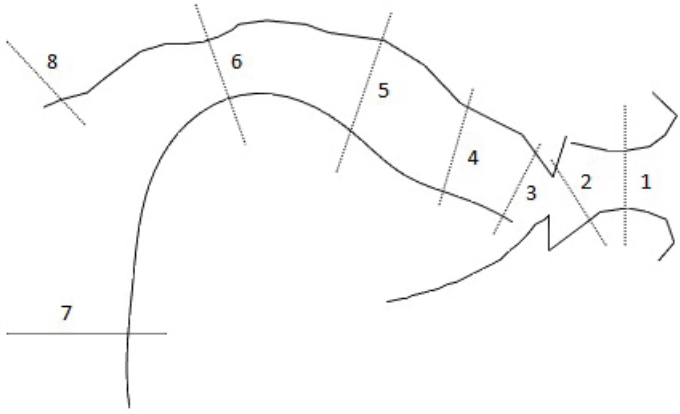

Рис. 5. Схема расположения 8-ми контрольных сечений

В качестве S_i, $i = 1, 2, \ldots, 8$ мы будем обозначать значение площади поперечного сечения в i-той области (например, S_1 — значение площади поперечного сечения на губах и т. д.). Введем еще dL — приращение длины речевого тракта относительно нейтральной длины (если тракт удлиняется относительно нейтрального состояния, то мы будем считать, что $dL > 0$; в противном случае (т. е. когда речевой тракт укорачивается) $dL < 0$). Для произвольного сечения произвольной фонемы площадь S_i будет либо удовлетворять ограничению-равенству ($S_i = a$, a — некоторое число), либо ограничению-неравенству ($b < S_i < c$; b, c — некоторые числа). Ограничения-равенства реализуются в первую очередь для смычных фонем (например, для фонем /Б, П/ имеем $S_1 = 0$). Подавляющее большинство фонем

[13] Авторы классических работ по динамическим моделям (например, [Saltzman, Munhall 1989]) особо отмечают, что целевая артикуляция фонемы в общем случае не совпадает с определенной конфигурацией речевого тракта. Именно по этой причине в рамках артикуляционной фонологии целевые артикуляции задаются на более глубинном уровне представления высказывания, чем соответствующие положения произносительных органов (см. далее).

[14] Предлагаемая схема контрольных областей условна и служит исключительно демонстрационным целям. С определенными оговорками эта схема подходит для русского литературного языка. Разумеется, для других языков число и схема расположения контрольных областей могут быть совершенно иными.

реализуют ограничения-неравенства (например, для фонем /С, З/ имеем $b_3 < S_3 < c_3$, где числа b_3, c_3 выбираются таким образом, чтобы площадь S_3 щели между кончиком языка и передними зубами оказывалась в диапазоне, необходимом для турбулизации воздушного потока, протекающего через эту щель).

Приведем примеры целевых артикуляций для некоторых фонем русского языка: 1) твердые смычные /П, Б/: $S_1 = 0$ (смычка на губах), $S_8 = 0$ (отсутствие назализации, противопоставляющее этим фонемам фонемы /М, М'/, $S_5 > a_5$ (отсутствие палатализации))[15], 2) фрикативные мягкие фонемы /С', З'/: $b_3 < S_3 < c_3$, $S_5 < a_5$, 3) твердая назальная фонема /Н/: $S_3 = 0$, $S_8 > 0$, $S_5 > a_5$, 4) гласная фонема /У/: $b_6 < S_6 < c_6$, $S_5 > a_5$, $S_7 > a_7$ (три ограничения определяют наличие необходимого сужения между средней частью языка и мягким небом при больших значениях площади поперечного сечения в передней и задней областях речевого тракта), $dL > thr$ (приращение длины речевого тракта должно превосходить некоторый порог thr)[16].

Из приведенных примеров ясно, что целевую артикуляцию можно определить еще так: «целевая артикуляция» — набор ограничений (равенств и неравенств) для данной фонемы, накладываемых на площади поперечного сечения в определенных контрольных областях речевого тракта, а также на приращение длины тракта. Такое определение дает возможность построить конструктивный алгоритм перехода от цепочки фонем, характеризующихся своими целевыми артикуляциями, к непрерывному артикуляционному потоку (см. ниже).

Каким образом устанавливаются целевые артикуляции фонем в том или ином языке? Для определения контрольных областей в данном языке, а также конкретных численных значений коэффициентов в ограничениях-равенствах и неравенствах используются специальные базы данных, содержащие измерения артикуляционных движений, совершаемых носителями этого языка при произнесении различного речевого материала (например, уже упомянутая нами база микролучевых рентгенографических измерений [Westbury 1994]). Проводя статистический анализ этих данных, можно определить количество и расположение контрольных областей, необходимых для порождения артикуляций по дискретной цепочке фонем в данном языке, а также подобрать нужные коэффициенты в ограничениях [Kaburagi, Honda 1996; Dusan 2000: 37–59; Okadome, Honda 2001]. Альтернативным подходом является подбор целевых артикуляций с помощью алгоритмов артикуляционного синтеза [Schroeter, Sondhi 1991]. Подбор коэффициентов в ограничениях-неравенствах для фрикативных фонем может быть выполнен как с помощью статистического анализа измеренных артикуляций, так и путем решения задачи рассеяния воздушного потока на препятствии, создаваемом губами, зубами или небом [Narayanan, Alwan 2000].

Понятие целевой артикуляции предполагает задание ограничений на площади сечения в определенных областях, но **не конкретизацию** того, с помощью каких артикуляций эти ограничения достигаются. Это свойство является ключевым для объяснения различных компенсационных эффектов в речевом потоке. Например, смычка ($S_3 = 0$) между кончиком языка и передними верхними зубами (как при артикуляции фонем /Т, Д/) может достигаться за счет использования одной из трех артикуляционных тактик: а) активное поднятие кончика языка к зубам при неизменном положении нижней челюсти, б) активное закрывающее движение нижней челюсти при неподвижном поднятом кончике языка, в) совместное использование активных движений кончика языка и нижней челюсти. Какая конкретная тактика будет использована человеком в данный момент времени — определяется

[15] Между собой эти фонемы различаются наличием / отсутствием голосового возбуждения, т. е. фонационным, а не артикуляционным признаком. Для поставленной задачи синтеза артикуляционных траекторий фонационные различия нерелевантны.

[16] Коэффициенты в ограничениях-равенствах и неравенствах для различных фонем должны быть различными. Приводимые в тексте целевые артикуляции не претендуют на полноту и преследуют демонстрационную цель.

текущими (в частности, экстралингвистическими) условиями произношения. Например, если в данный момент времени у человека практически обездвижена нижняя челюсть (ситуация сигареты в зубах или стоматологической анестезии), то будет использована тактика (а). В иных условиях человек может воспользоваться другими артикуляционными сценариями (никак не затрагивающими целевую артикуляцию переднеязычных зубных смычных фонем). Еще пример. На рис. 6 показаны две конфигурации речевого тракта при артикуляции фонемы /У/. Для данной фонемы в целевую артикуляцию входит условие удлинения речевого тракта относительно нейтрального состояния ($dL > thr$), при этом — на уровне целевой артикуляции — **не конкретизируется**, с помощью каких артикуляционных движений достигается это удлинение. На левом рисунке удлинение тракта достигается за счет лабиализации; на правом лабиализация отсутствует, а требуемое удлинение достигается за счет опускания гортани и уменьшения значения площади S_6 в области мягкого неба. Из рис. 6 видно, что для обеих конфигураций значения первых трех формантных частот очень близки. Какую конкретно артикуляционную тактику будет использовать человек для достижения требуемого удлинения — зависит от текущих условий произношения (например, в ситуации обездвиженных губ будет использована вторая тактика). В любом случае, соответствующая целевая артикуляция остается неизменной.

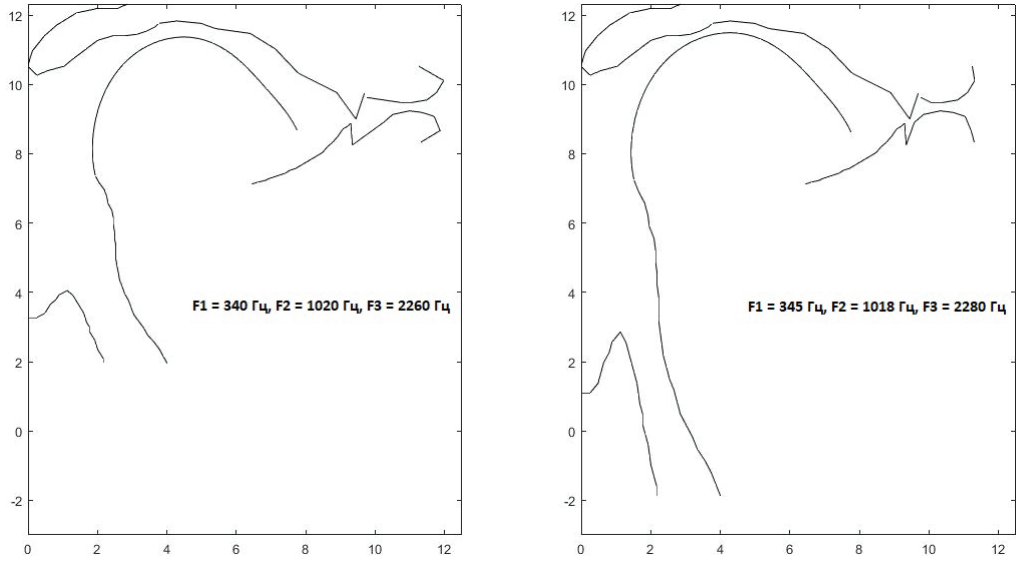

Рис. 6. Две артикуляционные конфигурации фонемы /У/ и соответствующие значения трех первых формантных частот. Необходимые значения формантных частот на левом графике достигаются за счет лабиализации, а на правом — за счет опускания гортани и уменьшения значения площади в области мягкого неба

Целевая артикуляция произвольной фонемы предполагает задание значений площадей лишь в определенных областях (своих для каждой фонемы) — в других же областях значения площадей могут лежать в весьма широких диапазонах значений. Это свойство является ключевым для объяснения очень многих коартикуляционных явлений в речевом потоке. Например, в целевую артикуляцию русских гласных фонем не входит площадь S_8 прохода в носовую область; соответственно, угол поворота небной занавески при их артикуляции лежит в весьма широких диапазонах. Конкретное значение угла поворота при произнесении русских гласных фонем будет определяться текущей фонетической позицией: в соседстве назальных звуков площадь прохода в носовой тракт будет ненулевой (произойдет

назализация гласных), а в соседстве неназальных звуков — нулевой (назализации не будет). Целевые артикуляции русских гласных фонем при этом останутся неизменными.

Понятие целевой артикуляции является ключевым для моделирования индивидуальных артикуляционных тактик. На рис. 7 (данные взяты из базы [Westbury 1994]) показана артикуляция звукосочетания /ARA/ двумя различными дикторами (три верхних рисунка соответствуют артикуляции звукосочетания /ARA/ диктором-мужчиной, три нижних рисунка — диктором-женщиной. Левый столбец соответствует артикуляции начального /A/, средний столбец — артикуляции /R/, правый столбец — артикуляции конечного /A/. Во всех случаях показана фаза выдержки для каждой из трех фонем. Временной интервал между конфигурациями на рисунках составляет примерно 150 мсек. В целевую артикуляцию фонемы /R/ входит ограничение-неравенство на значение площади в области между передней частью языка и твердым небом [Zhou et al. 2008]. Видно, что дикторы используют принципиально различные артикуляционные тактики достижения этой цели: в первом случае (средний столбец, верхний рисунок) цель достигается за счет подъема кончика языка к твердому небу и значительного изгиба кончика назад; во втором случае (средний столбец, нижний рисунок) цель достигается за счет подъема тела языка к твердому небу при опущенном кончике языка.

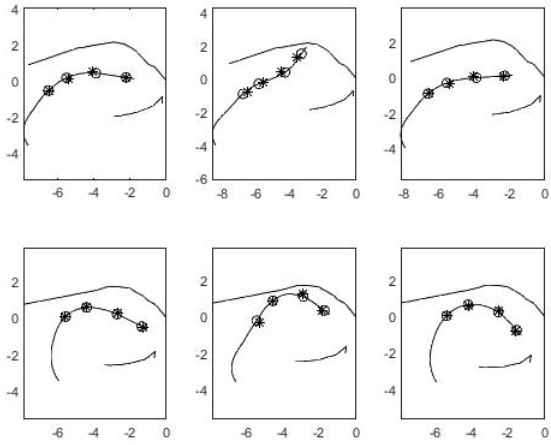

Рис. 7. Несколько фреймов для звукосочетания /ARA/. Верхняя строка — диктор-мужчина, нижняя — диктор-женщина. * — измеренные реперные точки на поверхности языка, o — соответствующие точки, вычисленные по модели. Временной интервал между кадрами примерно равен 150 мсек. На всех рисунках губы не показаны

Принципиально важным и до сих пор нерешенным вопросом является вопрос о том, насколько индивидуальное анатомическое строение речевых трактов людей влияет на определение целевых артикуляций тех или иных фонем. На рис. 8 показаны измеренные формы твердого неба для трех дикторов из базы [Westbury 1994]. Видно, что эти формы существенно различаются между собой. Отсюда следует, что и положения контрольных областей, и значения площадей в этих областях для разных дикторов могут быть различными. Как учесть такие различия в динамической фонологической модели? Нужно ли описывать каждую целевую артикуляцию с помощью некоторой вероятностной модели (например, для каждой целевой артикуляции строить модель гауссовых смесей, крайне популярную в системах распознавания речи [Huang et al. 2001])? Или же можно построить некоторый идеальный речевой тракт, в котором нивелированы все возможные анатомические различия между говорящими, определить целевые артикуляции для всех фонем относительно данного идеального тракта, а потом — при переходе к речевому тракту конкретного

человека — каким-то образом пересчитывать цели с учетом индивидуальных анатомических особенностей? Некоторые подходы к решению этой задачи предложены, например, в [Hashi et al. 1998; Serrurier et al. 2017; 2023], однако проблема все еще далека от окончательного решения. В рамках существующих динамических моделей синтез речевого потока обычно осуществляется для заданного речевого тракта с заранее известными анатомическими характеристиками (т. е. по сути для одного диктора). Именно такой подход принят в настоящем обзоре.

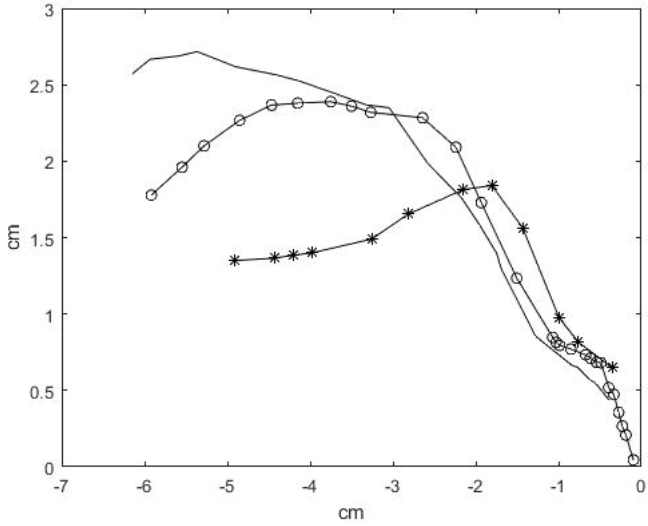

Рис. 8. Измеренные формы твердого неба для трех дикторов из базы [Westbury 1994]. В базе горизонтальная координата твердого неба отсчитывается от основания верхних резцов до точки прикрепления небной занавески, при этом значения этой координаты берутся со знаком «минус»

Итак, каждая фонема задается набором определенных значений площадей в некоторых контрольных сечениях речевого тракта. Естественно определить **фонологический артикуляционный признак** как площадь тракта в определенном контрольном сечении (в работах специалистов Хаскинских лабораторий артикуляционные признаки называются **phoneme tube features**, от слова *tube* — речевая труба). Например, губной признак соответствует площади S_1 в губной области, апикальный признак — площади S_3 в переднеязычно-зубной области и т. д. Признак — это переменная величина; для различных фонем он будет принимать различные **целевые** значения. Например, для фонем типа /Т, Д/ целевым значением переднеязычно-зубного признака будет ноль, а для фонем типа /С, З/ целевое значение того же признака должно оказаться внутри диапазона $b_3 < S_3 < c_3$ (b_3, c_3 — минимальное и максимальное значение площади, необходимой для турбулизации воздушного потока). Общее количество фонологических артикуляционных признаков в языке соответствует числу контрольных сечений; например, для схемы, показанной на рис. 5, число признаков = 8[17]. Определение термина «фонема» можно теперь ввести через термин «фонологический артикуляционный признак»: фонема — это набор целевых значений фонологических артикуляционных (и фонационных) признаков. Разные фонемы различаются а) наборами признаков, б) целевыми значениями признаков.

[17] Разумеется, для полной характеристики системы фонем в том или ином языке необходимо, помимо артикуляционных, указывать еще и фонационные признаки (а возможно, и что-то еще).

Представляет значительный интерес исследование того, как признаки меняют свои значения во времени (например, как меняется площадь губной области в процессе артикуляции определенного речевого материала). В рамках **артикуляционной фонологии** вводится специальный термин — «артикуляционный жест». Под артикуляционным жестом понимается любая траектория площади тракта $S_i(t)$ в i-том контрольном сечении (t — время), которая начинается с произвольного значения этой площади в некоторый начальный момент наблюдения t_0 и заканчивается целевым значением этой площади в заданный момент времени t_1. Мы уже видели, что одна и та же площадь в данном контрольном сечении может быть достигнута с помощью различных движений артикуляционных органов. Поэтому термин «артикуляционный жест» можно определить еще так: артикуляционный жест — набор координированных движений артикуляционных органов, осуществляемых для достижения целевого значения площади речевого тракта в определенном контрольном сечении в заданный момент времени[18].

При порождении конкретного речевого потока для каждого артикуляционного жеста необходимо задать набор **интервалов активации**, т. е. набор временных интервалов, содержащих следующую информацию: а) в какие моменты времени признак, соответствующий данному жесту, должен принимать определенные целевые значения (зависящие от входной цепочки фонем), б) сколько по времени должно выдерживаться то или иное целевое значение. Очень наглядно это можно отобразить с помощью **партитуры**, т. е. с помощью графика артикуляционного планирования данного высказывания. На рис. 9 показана партитура английской словоформы [bæn] (ban) для четырех артикуляционных жестов: губно-губного, палатального, апикального и назального.

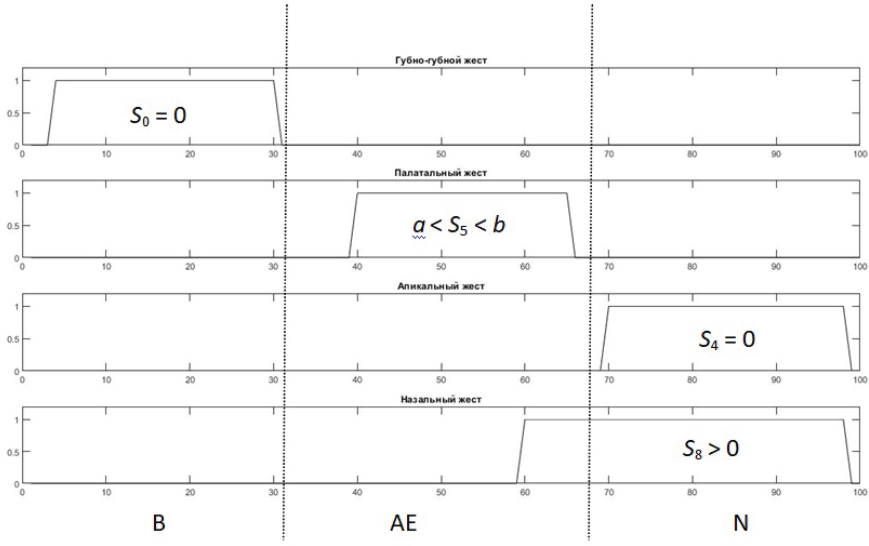

Рис. 9. Артикуляционная партитура словоформы [bæn] (ban).

По горизонтальной оси отложено время (в некоторых условных единицах); по вертикальной — амплитуда активации для каждого жеста. На временных интервалах, на которых амплитуда активации равна единице (на этих интервалах соответствующий жест

[18] Это классическое определение термина «артикуляционный жест» [Saltzman, Munhall 1989]. Иные определения артикуляционного жеста (например, жест — это некоторая артикуляция, жест — это звукотип и т. д.) не совпадают с классическим пониманием этого термина, что особо подчеркивалось в указанной работе.

считается активированным), площадь в соответствующем сечении должна быть равной целевому значению (например, на участке, соответствующем артикуляции фонемы /B/, площадь S_0 между губами должна быть нулевой; на участке, соответствующем фонеме /N/, должна быть достигнута смычка в переднеязычной области ($S_4 = 0$) и, кроме того, должна присутствовать назализация ($S_8 > 0$)). На интервалах, не являющихся интервалами активации, площади сечения соответствующих жестов могут быть практически произвольными (с условием, что они не будут нулевыми, а также не будут принимать слишком малые значения, достаточные для турбулизации воздушного потока в данном сечении). Какие конкретные значения примут площади соответствующих жестов на интервалах, не являющихся интервалами активации, будет определяться текущими артикуляционными ограничениями, а также условиями, задаваемыми принципом экономии произносительных усилий (см. ниже).

На этом мы заканчиваем обсуждение понятия **артикуляторный жест**. Мы еще к нему вернемся ниже, когда речь пойдет о работах исследователей из Хаскинских лабораторий в области артикуляторной фонологии. Сейчас только отметим, что не все динамические фонологические модели оперируют термином «артикуляторный жест»; например, в работах многих японских исследователей он не используется.

Итак, каждая фонема в данном языке представлена одной целевой артикуляцией. Имея на входе цепочку фонем, динамическая модель для каждой фонемы считывает соответствующую целевую артикуляцию, после чего по последовательности целей вычисляет траектории управляющих артикуляционных параметров. Для заданной артикуляционной модели эти траектории однозначно определяют последовательность конфигураций речевого тракта. Для понимания того, как именно происходит процесс перехода от последовательности целевых артикуляций к траекториям управляющих параметров, нам необходимы два важных термина: «критерий оптимальности» и «артикуляционные ограничения» (не путать с ограничениями, входящими в определение той или иной целевой артикуляции).

Под **артикуляционными ограничениями** понимаются любые ограничения (равенства или неравенства), накладываемые на возможные значения управляющих артикуляционных параметров. Эти ограничения разделяются на **глобальные** и **ситуативные**. Под глобальными понимаются ограничения на управляющие параметры, диктуемые физиологическими и физическими свойствами речевого аппарата. Например, угол поворота нижней челюсти (= угол раствора рта) лежит в определенных физиологически допустимых пределах и обычно не превышает 30°. Вертикальные и горизонтальные смещения языка, нижней челюсти и губ, а также вертикальное смещение гортани не могут превышать нескольких сантиметров. В силу механических и кинематических свойств артикуляционных органов допустимые скорости и ускорения, развиваемые языком, челюстью и губами, также находятся в определенных пределах (например, мышцы языка не могут развивать усилия в десятки кг). Примерные значения глобальных ограничений для некоторой артикуляционной модели приведены в [Сорокин 2012: 106]. В качестве глобальных ограничений могут использоваться ограничения-равенства, которые диктуются математическими алгоритмами, моделирующими артикуляционные траектории. Весьма распространенной является модель, описывающая динамику каждого управляющего параметра как отклик гармонического осциллятора на (ступенчатую или линейную) команду управления; при этом масса, упругость и коэффициент потерь данного осциллятора соответствует массе, упругости и коэффициенту потерь соответствующего управляющего параметра [Kaburagi, Honda 1996; Сорокин 2012: 104–106]. Выбор того или иного модельного ограничения-равенства (или неиспользование никакого) определяется теоретическими представлениями, вкусом и опытом конкретного исследователя.

Под ситуативными артикуляционными ограничениями понимаются ограничения на управляющие параметры, накладываемые конкретными текущими (в частности, экстралингвистическими) условиями произношения. Например, если в данный момент времени у человека блокировано движение губ, то диапазон расстояний между верхней и нижней

губой сократится до нуля. Возможны различные ситуативные ограничения на значения скоростей и ускорений управляющих параметров (например, человек говорит на бегу, эмоционально возбужден и т. д.).

Задание глобальных и/или ситуативных ограничений на скорости или ускорения управляющих параметров позволяет моделировать явление недостижения целевых артикуляций тех или иных фонем в речевом потоке — т. н. явление **undershoot** (недострел).

Рассмотрим теперь понятие **критерия оптимальности**. Как следует из предыдущего изложения, достижение целевых артикуляций в потоке речи возможно принципиально различными артикуляционными движениями; общее количество потенциально возможных артикуляционных траекторий бесконечно. Критерий оптимальности (совместно с глобальными и ситуативными ограничениями) определяет, какая конкретно последовательность конфигураций речевого тракта будет порождена в текущих условиях произношения. Критерий оптимальности — это принцип экономии произносительных усилий: человек из всех возможных артикуляционных траекторий выбирает такие, которые обеспечивают для него минимум затрачиваемых произносительных усилий[19].

С математической точки зрения[20] под **критерием оптимальности** понимается алгоритм, который по произвольной артикуляционной траектории управляющего параметра (или по целому набору произвольных траекторий различных управляющих параметров) определяет некоторое число. Простейший пример критерия оптимальности — среднее значение управляющего параметра на некотором интервале времени (мы берем траекторию управляющего параметра на некотором интервале времени и вычисляем среднее значение этого параметра на заданном интервале). Можно придумать бесконечное количество критериев оптимальности. В динамических фонологических моделях наиболее популярными являются: **энергетический** критерий (определяемый как сумма квадратов скоростей управляющего параметра в различные моменты времени) и **силовой** критерий (определяемый как сумма квадратов ускорений управляющего параметра в различные моменты времени) [Kaburagi, Honda 1996; Сорокин 2012: 335–372]. Энергетический критерий определяет суммарную кинетическую энергию, которую затрачивает человек в процессе артикуляции; силовой критерий есть суммарное усилие, которое человек затрачивает на осуществление движений артикуляционных органов. С фонетической точки зрения оба критерия могут быть интерпретированы как мера произносительных усилий, затрачиваемых человеком в процессе артикуляции. Отсюда следует, что само по себе словосочетание «произносительное усилие» не обозначает никакого физического или физиологического явления; однако смысл этого словосочетания вполне определяется, когда мы задаем конкретный критерий оптимальности[21].

В динамических фонологических моделях предполагается, что при планировании артикуляционных движений из всех возможных траекторий управляющих параметров выбираются только такие траектории, которые доставляют минимальное значение выбранному критерию оптимальности. Эта формулировка конкретизирует смысл словосочетания «принцип экономии произносительных усилий». Например, если в качестве критерия оптимальности используется энергетический критерий, то принцип экономии произносительных усилий означает следующее: из всех возможных артикуляционных траекторий

[19] Ср. [Кодзасов, Кривнова 2001: 88]: «Предполагается, что при планировании траекторий важную роль играет принцип экономии произносительных усилий».

[20] Дальнейшее описание в текущем абзаце вынужденно упрощенное.

[21] Энергетический и силовой критерии — не единственные критерии, используемые в динамических фонологических моделях. Весьма популярным является критерий работы на перемещение артикуляционных органов из нейтрального состояния (понимаемый как сумма квадратов смещений управляющего параметра из нейтрального состояния в различные моменты времени) [Сорокин 1992: 266 и сл.]. Возможно использование и комбинации тех или иных критериев [Kaburagi, Honda 1996].

человек выбирает такие, для которых суммарные энергетические затраты, связанные с совершением артикуляционных движений, минимальны. Для силового критерия принцип экономии произносительных усилий требует выбора только таких траекторий, для которых суммарное усилие на осуществление движений артикуляционных органов минимально.

Справедлив вопрос: откуда следует, что человек что-то вообще экономит при планировании и осуществлении артикуляционных движений? Вопрос очень тонкий, и его обсуждение завело бы нас слишком далеко в сторону от предмета настоящего обзора. Здесь мы ограничимся следующими пояснениями. Теоретическая механика учит, что механические системы, изменяя свое положение во времени, из всех возможных траекторий выбирают только такие, которые обеспечивают минимум некоторого критерия оптимальности (такое утверждение носит название «принципа наименьшего действия», см. [Арнольд 1999: 49])[22]. Поскольку речевой аппарат с физической точки зрения также является механической системой, он должен подчиняться принципу наименьшего действия. Весь вопрос в выборе такого критерия оптимальности, который бы адекватно моделировал наблюдаемые артикуляционные траектории. Опыт показывает, что траектории, минимизирующие энергетический или силовой критерии, во многих случаях аппроксимируют наблюдаемые артикуляционные движения с очень высокой точностью. В качестве примера на рис. 10 показаны измеренные траектории 4-х реперных точек из микролучевой базы [Westbury 1994] при произнесении звукосочетания /IA/ (см. тж. рис. 2-3, где те же самые траектории отображены на фоне модельных конфигураций речевого тракта). Траектории отрисованы линией «*-». Поверх них отображены траектории, вычисленные путем минимизации силового критерия оптимальности (отрисованы линией «o-»). Видно, что измеренные

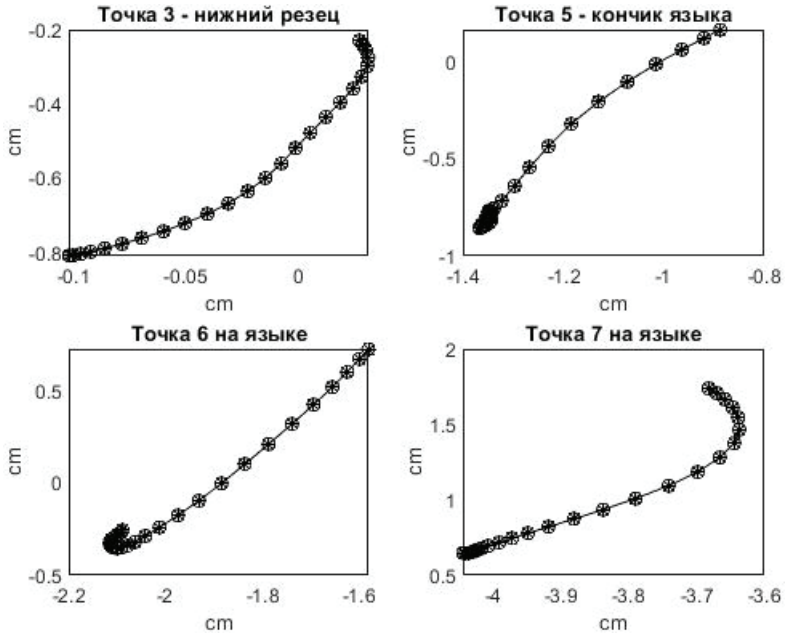

Рис. 10. Измеренные траектории четырех реперных точек при произнесении звукосочетания /IA/ (отображены как *-) и траектории, минимизирующие силовой критерий (отображены как o-)

[22] Отметим, что этот принцип не имеет рационального объяснения, что создает широкое поле для общефилософских и даже религиозных спекуляций.

и модельные траектории практически неотличимы друг от друга (по крайней мере, для данного речевого файла). Минимизация силового критерия оптимальности позволила в данном случае вычислить траектории, которые очень похожи на измеренные артикуляционные движения. Использование энергетического критерия также зачастую приводит к весьма точной аппроксимации измеренных траекторий.

Итак, для заданной последовательности целевых артикуляций в качестве соответствующих траекторий управляющих параметров выбираются такие, которые, во-первых, удовлетворяют всем глобальным и ситуативным ограничениям; во-вторых, обеспечивают экономию произносительных усилий; в-третьих, по возможности обеспечивают достижение целей в заданные моменты времени[23].

Вопрос о том, каким образом технически реализуется нахождение траекторий управляющих параметров, минимизирующих произносительные усилия, удовлетворяющих всем ограничениям и обеспечивающих достижение целевых артикуляций в заданные моменты времени, решается средствами теории оптимального управления, например методом динамического программирования или алгоритмами на базе принципа максимума Понтрягина. В силу значительных математических трудностей в настоящем обзоре этот вопрос не рассматривается (подробный алгоритм нахождения артикуляционных траекторий на базе динамического программирования представлен, например, в [Kaburagi, Honda 1996]).

Суммируем введенный понятийный аппарат в единую схему порождения речевого потока. Предполагается, что в памяти динамической фонологической модели означающие Лексикона записаны в виде цепочек фонем. Кроме того, в памяти модели хранится следующая информация: 1) целевая артикуляция для каждой фонемы, 2) артикуляционная модель, 3) список глобальных ограничений для управляющих параметров, 4) критерий оптимальности (или набор таких критериев).

На лингвистическом этапе построения высказывания из Лексикона извлекаются словоформы, которые организуются в некоторую синтаксическую структуру с помощью грамматических и синтаксических правил данного языка. Также на этом этапе осуществляются контекстные модификации означающих словоформ с помощью фонологических (лингвистических) правил — ассимиляции, редукции гласных, оглушения / озвончения конечных согласных и т. д. Лингвистический этап **не является** объектом изучения динамических фонологических моделей.

На вход динамической фонологической модели поступает некоторая цепочка фонем, полученная после завершения действия всех фонологических правил, и информация о текущих условиях, в том числе экстралингвистических, произношения[24]. Дальнейший синтез речевого потока удобно представить в виде следующей многоэтапной схемы.

Глубинно-моторный этап: на этом этапе для каждой фонемы из входной цепочки определяется соответствующая целевая артикуляция. Эквивалентным представлением является партитура артикуляционных жестов для входной цепочки фонем.

Этап внутреннего тайминга: на этом этапе расставляются моменты времени, в которые должны быть достигнуты целевые артикуляции. При расстановке временных меток

[23] Фраза «по возможности обеспечивают достижение целей в заданные моменты времени» означает следующее. Условия удовлетворения ограничениям и обеспечения экономии произносительных усилий выполняются всегда; напротив, условие достижения артикуляционных целей в речевом потоке для определенных фонем может и не выполняться (явление undershoot, недострел). Недострел произойдет в том случае, если при генерации траекторий управляющих параметров возникнет нарушение глобальных или ситуативных ограничений. Например, если для достижения цели в заданный момент времени необходимы такие скорости или ускорения некоторых управляющих параметров, которые физиологически или ситуативно недопустимы (= не удовлетворяют ограничениям), то в речевом потоке данная целевая артикуляция достигнута не будет.

[24] Вопрос о языке записи текущих условий произношения представляет большую практическую важность, однако, насколько автор может судить, совершенно не изучался в литературе.

учитываются как правила, свойственные данному языку, так и индивидуальные особенности организации темпо-ритмической структуры для данного человека (темп, свойственный данному человеку) и текущие условия произношения (текущее эмоциональное состояние человека, говорит ли он в положении покоя или на бегу и т. д.). Один из алгоритмов внутреннего тайминга построен в [Kaburagi, Kim 2007]. Если используется понятие партитуры, то этап внутреннего тайминга самостоятельно не выделяется — партитура уже содержит всю необходимую временную информацию для порождаемого высказывания.

Этап установки ситуативных ограничений: на этом этапе вводятся ситуативные ограничения на значения, скорости и ускорения управляющих параметров.

Поверхностно-моторный этап. На данном этапе происходит синтез артикуляционных конфигураций путем порождения таких траекторий, которые бы, во-первых, удовлетворяли всем глобальным и ситуативным ограничениям, во-вторых, минимизировали выбранный критерий оптимальности и, в-третьих, по возможности достигали бы целевых артикуляций в заданные моменты времени.

Далее мы рассмотрим две наиболее известные в мировой литературе динамические фонологические модели: 1) модели, построенные специалистами Хаскинских лабораторий, США (работы E. Salztman, K. Munhall, J. Perkell, D. Ostry, K. Browman, L. Goldstein и др.), 2) модели японских специалистов, в первую очередь сотрудников Waseda University и NTT Communication Science Labs (работы M. Honda, K. Honda, J. Dang, T. Kaburagi, T. Okadome и др.).

2. Специалисты Хаскинских лабораторий (Haskins Labs)

Основной фонологической концепцией, в рамках которой специалисты Хаскинских лабораторий описывают фонетику естественных языков, является так называемая **артикуляционная фонология** [Browman, Goldstein 1989]. В рамках данной концепции атомарной единицей фонологической системы для произвольного языка является артикуляционный жест[25]. Разные языки отличаются друг от друга различными наборами жестов и/или различными целевыми значениями артикуляционных признаков для тех или иных жестов. Моторная программа каждого планируемого высказывания записывается в виде партитуры (см. выше), на которой в явном виде указываются как интервалы активации для каждого жеста, так и целевые значения соответствующих артикуляционных признаков.

Понятие партитуры дает возможность ввести в фонологическое описание естественных языков время. Каждое планируемое высказывание на моторном уровне может быть представлено в виде той или иной партитуры; иначе говоря, на моторном уровне каждое высказывание является скоординированной во времени структурой определенных артикуляционных жестов. Для различных входных фонемных цепочек эти структуры различны; вместе с тем, в каждом языке присутствует набор типичных (для данного языка) артикуляционных структур. В рамках артикуляционной фонологии такие структуры называются фонологическими созвездиями (constellations), или фонологическими молекулами (molecules). Одни созвездия соответствуют фонемам (в традиционном таксономическом понимании этого термина), другие — слогам и т. д. Важно иметь в виду, что понятие «фонема» не является, в противоположность артикуляционному жесту, основополагающим в артикуляционной фонологии.

Интервалы активации для различных жестов могут перекрываться во времени — полностью или частично (например, на рис. 9 интервалы активации для палатального

[25] В работе [Byrd, Krivokapic 2021] предпринята попытка построить динамическую просодическую модель на базе постулатов артикуляционной фонологии. В рамках данной концепции атомарной единицей является просодический жест.

и назального жеста частично перекрываются во времени; на том же рисунке интервалы назального и апикального жестов полностью перекрываются во времени). В рамках артикуляционной фонологии явление временного перекрытия интервалов активации является основным для объяснения различных коартикуляционных явлений (например, частичной назализации гласного /æ/, очевидной на партитуре рис. 9). Различие между полным и неполным стилями произношения в рамках артикуляционной фонологии объясняется а) сокращением интервалов активации в неполном стиле, б) увеличением случаев частичного или полного перекрытия интервалов активации в неполном стиле, в) уменьшением амплитуды активаций для разных жестов (что означает недостижение теми или иными признаками своих целевых значений, undershoot) (подробнее см. [Browman, Goldstein 1989]).

Вычислительным модулем, позволяющим в рамках артикуляционной фонологии генерировать конкретные артикуляционные траектории по заданной дискретной цепочке фонем, является так называемая **task-dynamics**-модель, построенная в [Saltzman, Munhall 1989]. В рамках статьи порождение произвольного речевого потока по входной цепочке фонем разбивается на два этапа — глубинно-моторный этап («intergestural coordination», согласно терминологии [Ibid.]) и поверхностно-моторный этап («interarticulatory coordination», согласно терминологии [Ibid.]). На глубинно-моторном этапе осуществляется построение партитуры планируемого высказывания с помощью ряда эвристических правил. На поверхностно-моторном этапе происходит синтез конкретных артикуляционных траекторий.

В основе всех построений лежит артикуляционная модель, предложенная в [Rubin et al. 1981]. Модель характеризуется следующими управляющими параметрами («articulatory variables», согласно терминологии [Saltzman, Munhall 1989]): LH — горизонтальное смещение губ, JA — угол раствора нижней челюсти, ULV — вертикальное смещение верхней губы, LLV — вертикальное смещение нижней губы, TBR, TBA — вертикальное и горизонтальное смещение тела языка относительно системы координат, жестко связанной с нижней челюстью, TTR, TTA — вертикальное и горизонтальное смещение кончика языка, V — угол поворота небной занавески относительно точки ее прикрепления к твердому небу, G — расстояние между черпаловидными хрящами (т. е. фонационный признак, определяющий наличие / отсутствие голосового источника), т. е. всего 10 управляющих параметров. Кроме того, вводятся следующие 9 артикуляционных признаков («tract variables», согласно терминологии [Saltzman, Munhall 1989]), определяющих целевые артикуляции фонем: LP — степень огубления, LA — площадь поперечного сечения губной секции, LTH — площадь поперечного сечения зубной секции, TTCD — площадь поперечного сечения переднеязычной секции, TTCL — координата переднеязычной секции, TDCD — площадь поперечного сечения, соответствующего дорсальной области на языке, TDCL — координата соответствующей секции, VEL — площадь прохода в назальную область, GLO — площадь просвета между голосовыми складками. Артикуляционные признаки и управляющие параметры связаны набором линейных уравнений, полученных по результатам экспериментов с артикуляционной моделью из [Rubin et al. 1981].

С каждым из девяти артикуляционных признаков и каждой фонемой в данном языке связывается определенный артикуляционный жест. Каждый жест определяется четырьмя характеристиками: а) целевое значение артикуляционного признака для данной фонемы, б) эквивалентная масса секции речевого тракта, связанной с данным жестом (например, для билабиального жеста задается масса верхней и нижней губы), в) эквивалентная жесткость[26] секции речевого тракта, связанной с данным жестом, г) эквивалентный коэффициент механических потерь секции речевого тракта, связанной с данным жестом. Динамика каждого жеста описывается уравнением гармонического осциллятора[27]. В качестве

[26] Жесткость некоторого речевого участка является мерой его напряженности: чем выше жесткость, тем больше напряжены стенки речевого тракта на данном участке, и наоборот.

[27] Читатели, не знакомые с уравнением гармонического осциллятора, могут обратиться за дополнительной информацией к любому учебнику по обыкновенным дифференциальным уравнениям.

критерия оптимальности используется критерий минимума работы управляющих параметров на смещение относительно нейтральной конфигурации речевого тракта. Построение интервалов активации для каждого жеста осуществляется с помощью специальных лингвистических правил; эти правила совместно с критерием минимума работы артикуляторов, уравнениями гармонического осциллятора и аналитическими соотношениями между артикуляционными признаками и управляющими параметрами полностью определяют динамическую модель [Saltzman, Munhall 1989]. Выстраивая для каждой входной цепочки фонем соответствующую партитуру и используя упомянутые выше уравнения, можно вычислить искомый речевой поток.

Используя построенную динамическую модель, авторы провели ряд экспериментов по моделированию некоторых артикуляционных явлений:

1. Были смоделированы звукосочетания /i g i/ и /æ g æ/, при этом конфигурации начальной и конечной гласной фонемы были заданы. Для согласной фонемы /g/ в качестве целевой артикуляции задавалась только нулевая велярная смычка, но не место смыкания (место смыкания должно было быть рассчитано динамической моделью). Для /g/ в окружении гласных /i/ вычисленное место смыкания справедливо оказалось гораздо более передним (палатализованным), чем для /g/ в окружении гласных /æ/.
2. Были сгенерированы траектории нескольких реперных точек на губах, нижней челюсти и языке при артикуляции звукосочетания /$\Gamma_1 — C_1 — \Gamma_2 — C_2 — \Gamma_3$/, где в качестве C_1 и C_2 выступали глухие и звонкие билабиальные взрывные фонемы, а в качестве Γ_1, Γ_2 и Γ_3 выступали гласные фонемы /ə, æ, ə/. Вычисленные траектории были сопоставлены с траекториями, измеренными с помощью скоростной кинорентгеносъемки для некоторого носителя американского английского языка. Траектории качественно оказались весьма похожими; к сожалению, количественные результаты сравнения траекторий в статье не приводятся.

В [Browman, Goldstein 1990] артикуляционная партитура дополнена информацией о ритмической организации порождаемого высказывания. Кроме того, показана необходимость использования более глубинных, чем партитура, представлений высказывания в терминах гласных и согласных классов фонем. Полученные результаты обсуждаются применительно к проблеме моделирования речевого потока в неполном стиле произношения.

Построение партитуры для произвольного высказывания требует задания интервалов активации артикуляционных жестов, а также временных сдвигов этих интервалов друг относительно друга (intergestural timing). Правила построения таких сдвигов для полного и неполного стилей произношения на материале различных языков обсуждаются в [Browman 1992; Gafos et al. 2020]. Более сложная модель, опирающаяся на методы нелинейной динамики, построена и исследована в [Saltzman, Byrd 2000].

В работах [Smith et al. 1993; McGowan 1994; McGowan, Lee 1996] построены алгоритмы вычисления характеристик артикуляционных жестов (т. е. значений эквивалентных масс, коэффициентов потерь и жесткостей) по измеренным траекториям реперных точек на подвижных поверхностях речевого тракта и по формантным траекториям.

В [McGowan, Saltzman 1995] в число управляющих параметров артикуляционной модели были внесены три новых: а) сила, оказываемая диафрагмой на легкие, б) натяжение голосовых складок и в) объем речевого тракта от голосовой щели до места наибольшего сужения во рту. Состав признаков также был дополнен тремя новыми: а) подсвязочное давление, б) перепад надсвязочного и подсвязочного давления и в) частота основного тона. Новые управляющие параметры и признаки позволили авторами моделировать не только артикуляцию, но и некоторые аэродинамические процессы, происходящие в речевом тракте в процессе речеобразования. В статье по входным цепочкам фонем для некоторых звукосочетаний типа «гласная фонема + согласная фонема + гласная фонема» синтезированы не только речевые потоки (= последовательности артикуляций), но и соответствующие этим потокам акустические сигналы. В [Rubin et al. 1996] набор управляющих параметров

дополнен еще одним — параметром, определяющим прогиб кончика языка (TTCO) в средне-сагиттальном сечении. Кроме того, предложена модель автоматического построения интервалов активации на базе некоторой рекуррентной нейронной сети. Обновленная динамическая модель с успехом была использована в артикуляционном синтезаторе CASY [Iskarous et al. 2003; Nam et al. 2013].

3. Японские исследования (Waseda University, NTT Communication Science Labs)

Классической статьей, в значительной степени определившей направление работ японских исследователей, стала публикация [Kaburagi, Honda 1996]. В ней в качестве данных использовались синхронные измерения траекторий семи реперных точек (одна точка на верхней губе, одна на нижней, одна на нижнем резце, четыре точки на языке) и соответствующих речевых сигналов. Все измерения выполнены с помощью электромагнитного артикулографа. В качестве минимальной единицы Лексикона выступает фонема. Управляющими параметрами служили: расстояние между верхней и нижней губой, степень огубления, координаты нижнего резца и координаты 32-х управляющих точек на поверхности языка. Динамика каждого управляющего параметра моделировалась уравнением гармонического осциллятора, возбуждаемого ступенчатой командой (с физиологической точки зрения команда есть равнодействующая всех сил, оказываемых мышцами на данный управляющий параметр). В качестве критерия оптимальности использовалась сумма энергетического критерия и дополнительного критерия, минимизирующего количество команд на некотором временном интервале (около 150 мсек). Целевые артикуляции для каждой фонемы определялись посредством специальных ограничений-равенств, связывающих управляющие параметры с сужениями в различных сечениях речевого тракта. Полная процедура автоматического вычисления артикуляционных траекторий по заданной дискретной цепочке фонем сводилась к минимизации суммарного энергетического критерия при двух типах ограничений: а) ограничения-равенства, задаваемые целевыми артикуляциями фонем в определенные моменты времени, б) модельные ограничения-равенства, связывающие динамику каждого управляющего параметра с командой посредством уравнения гармонического осциллятора. В статье построен алгоритм аналитического решения этой оптимизационной задачи на базе динамического программирования.

В [Kaburagi, Honda 2001] разработана специальная статистическая процедура (на базе линейного дискриминантного анализа), позволяющая по измерениям с помощью электромагнитного артикулографа определять целевые артикуляции фонем (всего в исследовании использовалось 35 фонем японского языка). Используя данную процедуру, а также упомянутый выше алгоритм расчета речевого потока по дискретной цепочке фонем, авторы смоделировали ряд коартикуляционных эффектов в японском языке на синтетических траекториях некоторых звукосочетаний типа «гласная фонема + согласная фонема + гласная фонема» (гласные = /E, I/, согласные = /P, T, K/). На интервале произнесения всех согласных фонем алгоритм правильно автоматически рассчитал палатализацию языка перед /I/ и веляризацию перед /A/[28]. Также были смоделированы речевые потоки для ряда фраз японского языка. Траектории для каждой реперной точки были сопоставлены с траекториями, измеренными с помощью электромагнитного артикулографа. Средняя ошибка между измеренными и вычисленными траекториями по всем точкам составила 1,5 мм

[28] Целевые артикуляции для /P, T, K/ в этой работе определялись как нулевая площадь на губах, в переднеязычной и заднеязычной областях соответственно. Палатализация и веляризация не входили в число целевых артикуляций, поскольку оппозиция по «твердости — мягкости» в японском языке не является смыслоразличительной.

(к сожалению, значения среднеквадратической ошибки в процентах не приводятся). Также были проведены эксперименты по синтезу речевых потоков при условии, что минимальной единицей Лексикона является не фонема, а дифон (всего в исследовании использовалось 248 дифонов). Выяснилось, что точность аппроксимации измеренных траекторий фонемной и дифонной моделями практически одинакова. Отсюда был сделан вывод о физиологической адекватности построенной фонемной модели.

В [Okadome, Honda 2001] в качестве минимальной единицы Лексикона использовался трифон (всего использовалось 507 трифонов). В качестве целевой артикуляции каждого трифона использовались средние значения координат и скоростей управляющих параметров, определенных с помощью базы артикулографических измерений. В качестве целевого функционала использовался силовой критерий. Средняя ошибка между измеренными и вычисленными траекториями по всем реперным точкам составила 1,55 мм, что сопоставимо с точностью фонемной динамической модели.

В [Kaburagi, Kim 2007] исследовалась задача порождения как артикуляционных траекторий, так и соответствующих речевых сигналов по дискретной цепочке фонем. В качестве реперных точек использовались 8 точек измерения (7 точек на губах, нижнем резце и языке — см. выше, и одна точка на мягком небе). В качестве единицы Лексикона использовалась фонема. Артикуляторная модель, ограничения и критерий оптимальности заимствовались из [Kaburagi, Honda 1996]. Процедура определения целевых артикуляций фонем совпадала с процедурой из [Kaburagi, Honda 2001]. Построен алгоритм внутреннего таймирования. С помощью базы синхронных измерений речевых сигналов и соответствующих траекторий реперных точек каждой фонеме был сопоставлен соответствующий ей акустический спектр (акустические спектры параметризовались 14 коэффициентами линейного предсказания при частоте дискретизации речевого сигнала = 8 кГц). После генерации артикуляционных траекторий акустический спектр в произвольный момент времени вычислялся как взвешенная сумма акустических спектров, соответствующих двум соседним фонемам. Эксперименты с синтезом артикуляционных траекторий и соответствующих им сонограмм были проведены на материале ряда звукосочетаний японского языка типа «гласная фонема + согласная фонема + гласная фонема», а также на одной японской фразе. Синтетические траектории и соответствующие им сонограммы оказались качественно весьма похожими на артикуляционные и акустические измерения. Среднее значение среднеквадратических погрешностей между измеренными и синтетическими траекториями для всех фраз составила 1,27 мм, а среднее значение различий между измеренными спектрами и спектрами, вычисленными по артикуляциям, составило 3,44 дБ. Таким образом, полученные результаты можно считать обнадеживающими.

В упомянутых выше работах использовалась сравнительно простая артикуляторная модель. Гораздо более сложная модель построена в [Dang, Honda 2004]. В рамках модели тело языка в трех измерениях аппроксимируется набором связанных друг с другом упругих параллелепипедов. Девять управляющих язычных мышц (genioglossus, styloglossus, hyoglossus, verticalis и т. д.) описываются упругими цилиндрами. Управляющими параметрами служат амплитуды сил, действующих на язычные мышцы. Построен алгоритм, позволяющий пересчитывать мышечные усилия в сужения в тех или иных областях речевого тракта. Исследован вопрос о реорганизации партитуры мышечных усилий при генерации различных компенсационных артикуляционных явлений. В [Ito et al. 2004] предложено дополнить набор управляющих параметров значениями жесткости соответствующих язычных мышц.

В [Chen et al. 2013; Yan et al. 2014] построена статистическая динамическая фонологическая модель. В рамках модели отображение пространства мышечных усилий на целевые артикуляции и соответствующие управляющие параметры осуществляется с помощью специальных нейронных сетей (так называемых самоорганизующихся карт Кохонена). Построенная модель позволила сгенерировать ряд коартикуляционных явлений в различных сочетаниях «согласная фонема + гласная фонема» для японского языка.

Процедура определения целевых артикуляций, предложенная в [Kaburagi, Honda 2001], хорошо подходит для согласных фонем; для гласных же эта процедура гораздо менее эффективна. Работа [Lu, Dang 2010] посвящена определению целевых артикуляций гласных фонем японского языка методами алгебраической топологии. Построен алгоритм, позволяющий вычислять целевые артикуляции по базе измерений, выполненных с помощью электромагнитного артикулографа.

Ряд работ посвящен проблеме учета индивидуальных анатомических особенностей речевого тракта при определении целевых артикуляций. В [Hashi et al. 1998] вместо стандартной декартовой системы координат введены две новые координатные оси — одна соответствует расстоянию, измеренному вдоль поверхности мягкого и твердого неба (т. е. речевой тракт в этих областях как бы выпрямляется), а вторая ось откладывается перпендикулярно первой. Для некоторых гласных звуков и некоторых дикторов подобная нормализация позволила снизить дисперсию измерений реперных точек на поверхности языка на 1 мм (по сравнению с дисперсией исходных измерений). Другой подход предложен в [Wei, Dang 2008]. В рамках этого подхода соответствие между реперными точками произвольного диктора с произвольными анатомическими и артикуляторными характеристиками и реперными точками идеального речевого тракта строится с помощью формулы для тонкой бесконечной пластины. Дисперсия измерений всех реперных точек снизилась на 2 мм по сравнению с дисперсией исходных ненормированных измерений. При этом процедура нормализации практически не затронула расположение контрольных сечений, в которых определяются целевые значения площадей речевого тракта.

Заключение

На этом мы заканчиваем обзор основных научных результатов, полученных специалистами Хаскинских лабораторий и исследователями из Waseda University и NTT Communication Science Labs. Рамки обзора не позволили осветить ряд важных тем. Например, мы опустили вопрос о сопоставлении артикуляционной фонологии с другими фонологическими теориями (традиционная таксономическая фонология, генеративная фонология, автосегментная фонология и т. д.) (детально этот вопрос изложен в [Browman, Goldstein 1989; Studdart-Kennedy, Goldstein 2003; Honorof 2004]). Мы также не коснулись того, каким образом в рамках динамических моделей объясняются процессы усвоения фонетической структуры языка детьми (см. [Goldstein, Fowler 2003; Rubertus, Noiray 2018]). За рамками обзора осталась проблема поиска фонетических и фонологических коррелятов в человеческом мозгу [Mascheretti et al. 2021; Ohashi, Ostry 2021]. Все же автор надеется, что ему удалось в какой-то степени рассказать об основных положениях теории динамических моделей и, возможно, даже заинтересовать кого-то из читателей.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Арнольд 1999 — Арнольд В. И. *Лекции об уравнениях с частными производными*. М.: ФАЗИС, 1999. [Arnol'd V. I. *Lektsii ob uravneniyakh s chastnymi proizvodnymi* [Lectures on partial differential equations]. Moscow: FAZIS, 1999.]

Баден и др. 2005 — Баден П., Макаров И. С., Сорокин В. Н. Алгоритм вычисления площадей поперечных сечений речевого тракта. *Акустический журнал*, 2005, 51(1): 52–58. [Badin P., Makarov I. S., Sorokin V. N. An algorithm for calculating the cross-section areas of the vocal tract. *Akusticheskii zhurnal*, 2005, 51(1): 52–58.]

Князев 1999 — Князев С. В. О прогрессивной ассимиляции в современном русском языке. *Вестник Московского государственного университета. Сер. 9. Филология*, 1999, 4: 18–33. [Knyazev S. V.

On progressive assimilation in Modern Russian. *Vestnik Moskovskogo gosudarstvennogo universiteta. Ser. 9. Filologiya*, 1999, 4: 18–33.]

Князев 2001 — Князев С. В. Еще раз о соотношении фонетики и фонологии. *Вестник Московского государственного университета. Сер. 9. Филология*, 2001, 5: 101–112. [Knyazev S. V. Once again on the relationship between phonetics and phonology. *Vestnik Moskovskogo gosudarstvennogo universiteta. Ser. 9. Filologiya*, 2001, 5: 101–112.]

Князев 2004 — Князев С. В. Об иерархии фонологических правил в русском языке (несколько новых соображений по поводу *язв* А. А. Реформатского). *Семиотика, лингвистика, поэтика: К столетию со дня рождения А. А. Реформатского*. Виноградов В. А. (ред.). М.: Языки славянской культуры, 2004, 151–166. [Knyazev S. V. On the hierarchy of phonological rules in Russian (several new thoughts on *yazv* of A. A. Reformatsky). *Semiotika, lingvistika, poetika: K stoletiyu so dnya rozhdeniya A. A. Reformatskogo*. Vinogradov V. A. (ed.). Moscow: Yazyki slavyanskoi kul'tury, 2004, 151–166.]

Кодзасов, Кривнова 2001 — Кодзасов С. В., Кривнова О. Ф. *Общая фонетика: Учебник*. М.: РГГУ, 2001. [Kodzasov S. V., Krivnova O. F. *Obshchaya fonetika: Uchebnik* [General phonetics: A textbook]. Moscow: Russian State Univ. for the Humanities, 2001.]

Леонов и др. 2003 — Леонов А. С., Макаров И. С., Сорокин В. Н., Цыплихин А. И. Артикуляторный ресинтез гласных. *Информационные процессы*, 2003, 3(2): 73–82. [Leonov A. S., Makarov I. S., Sorokin V. N., Tsyplikhin A. I. Articulatory resynthesis of vowels. *Informatsionnye protsessy*, 2003, 3(2): 73–82.]

Леонов и др. 2004 — Леонов А. С., Макаров И. С., Сорокин В. Н., Цыплихин А. И. Артикуляторный ресинтез фрикативных. *Информационные процессы*, 2004, 4(2): 141–159. [Leonov A. S., Makarov I. S., Sorokin V. N., Tsyplikhin A. I. Articulatory resynthesis of fricatives. *Informatsionnye protsessy*, 2004, 4(2): 141–159.]

Леонов и др. 2005 — Леонов А. С., Макаров И. С., Сорокин В. Н., Цыплихин А. И. Кодовая книга для речевых обратных задач. *Информационные процессы*, 2005, 5(2): 101–119. [Leonov A. S., Makarov I. S., Sorokin V. N., Tsyplikhin A. I A codebook for speech inverse problems. *Informatsionnye protsessy*, 2005, 5(2): 101–119.]

Макаров 2005 — Макаров И. С. *Построение и исследование артикуляторных кодовых книг для решения речевых обратных задач*. Дис. … канд. тех. н. М.: ИППИ РАН, 2005. [Makarov I. S. *Postroenie i issledovanie artikulyatornykh kodovykh knig dlya resheniya rechevykh obratnykh zadach* [Construction and research of articulatory codebooks for solution of speech inverse problems]. Candidate diss. Moscow: Kharkevich Institute for Information Transmission Problems, 2005.]

Макаров 2009 — Макаров И. С. Аппроксимация речевого тракта коническими рупорами. *Акустический журнал*, 2009, 55(2): 256–265. [Makarov I. S. Approximating the vocal tract by conical horns. *Akusticheskii zhurnal*, 2009, 55(2): 261–269.]

Макаров 2011 — Макаров И. С. Алгоритм быстрого вычисления передаточной функции для неоднородной акустической трубы. *Акустический журнал*, 2011, 57(5): 695–708. [Makarov I. S. A fast transfer function algorithm for nonuniform acoustic tubes. *Akusticheskii zhurnal*, 2011, 57(5): 695–708.]

Макаров 2023 — Макаров И. С. Двухмерная математическая модель языка. Рук., 2023. [Makarov I. S. *Dvukhmernaya matematicheskaya model' yazyka* [A 2D-tongue mathematical model]. Ms., 2023.]

Макаров, Сорокин 2004 — Макаров И. С., Сорокин В. Н. Резонансы разветвленного речевого тракта с податливыми стенками. *Акустический журнал*, 2004, 50 (3): 389–396. [Makarov I. S., Sorokin V. N. Resonances of the branched vocal tract with compliant walls. *Akusticheskii zhurnal*, 2004, 50(3): 323–330.]

Сорокин 1992 — Сорокин В. Н. *Синтез речи*. М.: Наука, 1992. [Sorokin V. N. *Sintez rechi* [Speech Synthesis]. Moscow: Nauka, 1992.]

Сорокин 2012 — Сорокин В. Н. *Речевые процессы*. М.: Народное образование, 2012. [Sorokin V. N. *Rechevye protsessy* [Speech Processes]. Moscow: Narodnoe obrazovanie, 2012.]

Щерба 1974 — Щерба Л. В. О разных стилях произношения и об идеальном фонетическом составе слов. *Языковая система и речевая деятельность*. Зиндер Л. Р., Матусевич М. И. (ред.). Л.: Наука, 1974. [Shcherba L. V. On different pronunciation styles and on the ideal phonetic structure of words. *Yazykovaya sistema i rechevaya deyatel'nost'*. Zinder L. R., Matusevich M. I. (eds.). Leningrad: Nauka, 1974.]

Browman 1992 — Browman C. Articulatory Phonology: An overview. *Phonetica*, 1992, 49: 155–180.

Browman, Goldstein 1989 — Browman C., Goldstein L. Articulatory gestures as phonological units. *Phonology*, 1989, 6: 201–251.

Browman, Goldstein 1990 — Browman C., Goldstein L. Tiers in articulatory phonology, with some implications for casual speech. *Papers in laboratory phonology I: Between the grammar and the physics of speech*. Kingston J., Beckman M. E. (eds.). Cambridge: Cambridge Univ. Press, 1990, 341–376.

Byrd, Krivokapic 2021 — Byrd D., Krivokapic E. Cracking prosody in articulatory phonology. *Annual Review in Linguistics*, 2021, 7: 31–53.

Chen et al. 2013 — Chen X., Dang J., Yan H., Fang Q., Kröger B. A neural understanding of speech motor learning. *Proc. of the 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (Kaohsiung, Oct. 29–Nov. 1, 2013)*. New York: Institute of Electrical and Electronics Engineers, 2013, 1–14.

Dang, Honda 2004 — Dang J., Honda K. Construction and control of a physiological articulatory model. *Journal of the Acoustical Society of America*, 2004, 115(2): 853–870.

Dusan 2000 — Dusan S. *Statistical estimation of articulatory trajectories from the speech Signal using dynamic and phonological constraints*. Ph.D. diss. Waterloo: Univ. of Waterloo, 2000.

Fant 2001 — Fant G. Swedish vowels and a new three-parameter model. *Quarterly Progress and Status Report*, 2001, 42(1): 43–49.

Gafos et al. 2020 — Gafos A., Roeser J., Sotiropoulou S., Hoole P., Zeroual C. Structure in mind, structure in vocal tract. *Natural Language Linguistic Theory*, 2020, 38: 43–75.

Goldstein, Fowler 2003 — Goldstein L., Fowler C. Articulatory Phonology: A phonology for public language use. *Phonetics and phonology in language Comprehension and production*. Schiller N. O., Meyer A. S. (eds.). Berlin: Mouton de Gruyter, 2003, 159–207.

Gomez et al. 2020 — Gomez A., Stone M. L., Woo J., Xing F., Prince J. L. Analysis of fiber strain in the human tongue during speech. *Computer Methods in Biomechanics and Biomedical Engineering*, 2020: 23(8), 312–322.

Hanson, Stevens 2002 — Hanson H., Stevens K. A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn. *Journal of the Acoustical Society of America*, 2002, 112: 1158–1182.

Hashi et al. 1998 — Hashi M., Westbury J. R., Honda K. Vowel posture normalization. *Journal of the Acoustical Society of America*, 1998, 104: 2426–2437.

Honorof 2004 — Honorof D. Articulatory events are given in advance. *Hard-Science Linguistics*. Yngve V. H., Wasik Z. (eds.). London: Continuum, 2004, 67–86.

Huang et al. 2001 — Huang X., Acero A., Hon H.-W. *Spoken Language Processing*. New Jersey: Prentice Hall, 2001.

Iskarous et al. 2003 — Iskarous K., Goldstein L., Whalen D., Tiede M., Rubin P. CASY: The Haskins Configurable Articulatory Synthesizer. *Proc. of the 15th International Congress of Phonetic Sciences (Barcelona, Aug. 3–9, 2003)*. Solé M. J., Recasens D., Romero J. (eds.). Barcelona: FUTURGRAFIC, 2003, 185–188.

Ito et al. 2004 — Ito T., Gomi H., Honda M. Dynamical simulation of speech cooperative articulation by muscle linkages. *Biological Cybernetics*, 2004, 91: 275–282.

Kaburagi, Honda 1994 — Kaburagi T., Honda M. Determination of sagittal tongue shape from the positions of points on the tongue surface. *Journal of the Acoustical Society of America*, 1994, 96(3): 1356–1366.

Kaburagi, Honda 1996 — 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.

Kaburagi, Honda 2001 — Kaburagi T., Honda M. Dynamic articulatory model based on multidimensional invariant-feature task representation. *Journal of the Acoustical Society of America*, 2001, 110(1): 441–451.

Kaburagi, Kim 2007 — Kaburagi T., Kim J. Generation of the vocal tract spectrum from the underlying articulatory mechanism. *Journal of the Acoustical Society of America*, 2007, 121(1): 456–468.

Lu, Dang 2010 — Lu X., Dang J. Vowel Production Manifold: Intrinsic Factor Analysis of Vowel Articulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, 18(5): 1053–1062.

Mascheretti et al. 2021 — Mascheretti S., Perdue M., Feng B., Andreola Ch., Dionne G., Jasińska K., Pugh K., Grigorenko E., Landi N. From BDNF to reading: Neural activation and phonological processing as multiple mediators. *Behavioral Brain Research*, 2021, 396: 112859.

McGowan 1994 — McGowan R. Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication*, 1994, 14: 19–48.

McGowan, Lee 1996 — McGowan R., Lee M. Task dynamic and articulatory recovery of lip and velar approximations under model mismatch conditions. *Journal of the Acoustical Society of America*, 1996, 99(1): 595–608.

McGowan, Saltzman 1995 — McGowan R., Saltzman E. Incorporating aerodynamic and laryngeal components into task dynamics. *Journal of Phonetics*, 1995, 23: 255–269.

Nam et al. 2013 — Nam H., Mooshammer Ch., Iskarous K., Whalen D. Hearing tongue loops: Perceptual sensitivity to acoustic signatures of articulatory dynamics. *Journal of the Acoustical Society of America*, 2013, 134(5): 3808–3817.

Narayanan, Alwan 2000 — Narayanan S., Alwan A. Noise source models for fricative consonants. *IEEE Transactions on Speech and Audio Processing*, 2000, 8(3): 328–344.

Ohashi, Ostry 2021 — Ohashi H., Ostry D. Neural Development of Speech Sensorimotor Learning. *The Journal of Neuroscience*, 2021, 41(18): 4023–4035.

Okadome, Honda 2001 — Okadome T., Honda M. Generation of articulatory movements by using a kinematic triphone model. *Journal of the Acoustical Society of America*, 2001, 110(1): 453–462.

Rubertus, Noiray 2018 — Rubertus E., Noiray A. On the development of gestural organization: A cross-sectional study of vowel-to-vowel anticipatory coarticulation. *PLOS ONE*, 2018, 13(9): 1–21.

Rubin et al. 1981 — Rubin P., Baer T., Mermelstein P. An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 1981, 70, 321–328.

Rubin et al. 1996 — Rubin P., Saltzman E., Goldstein L., McGowan R., Tiede M., Browman C. CASY and extensions to the task-dynamic model. *1st ESCA Tutorial and Research Workshop on Speech Prodution Modeling — 4th Speech Production Seminar (Autrans, May 20–24, 1996)*. 1996, 125–128.

Saltzman, Byrd 2000 — Saltzman E., Byrd D. Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, 2000, 19: 499–526.

Saltzman, Munhall 1989 — Saltzman E., Munhall K. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1989, 1(4): 333–382.

Schroeter, Sondhi 1991 — Schroeter J., Sondhi M. M. Speech coding based on physiological models of speech production. *Advances in speech signal processing*. Furui S., Sondhi M. M. (eds.). New York: Marcel Dekker, 1991, 231–266.

Serrurier et al. 2017 — Serrurier A., Badin P., Boe L.-J., Lamalle L., Neuschaefer-Rube C. Inter-speaker variability: Speaker normalisation and quantitative estimation of articulatory invariants in speech production for French. *Proc. of 18th Annual Conf. of the International Speech Communication Association (Stockholm, Aug. 20–24, 2017).* Red Hook (NY): Curran Associates, 2017: 2272–2276.

Serrurier et al. 2023 — Serrurier A., Neuschaefer-Rube Ch. Morphological and acoustic modeling of the vocal tract. *Journal of the Acoustical Society of America*, 2023, 153: 1867–1886.

Smith et al. 1993 — Smith C., Browman C., McGowan R., Kay B. Extracting dynamic parameters from speech movement data. *Journal of the Acoustical Society of America*, 1993, 93(3): 1580–1586.

Sorokin et al. 2005 — Sorokin V. N., Leonov A. S., Makarov I. S., Tsyplikhin A. I. Speech inversion and resynthesis. *Proc. of the 6th Interspeech 2005 and 9th European Conf. on Speech Communication and Technology (Lisboa, Sept. 4–8, 2005)*. Red Hook (NY): Curran Associates, 2005, 3209–3212.

Story, Bunton 2011 — Story B., Bunton K. Decomposition of vowel and consonant contributions to the time-varying vocal tract shape. *Journal of the Acoustical Society of America*, 2011, 129, 2456.

Story, Bunton 2019 — Story B., Bunton K. A model of speech production based on the acoustic relativity of the vocal tract. *Journal of the Acoustical Society of America*, 2019, 146: 2522–2528.

Studdart-Kennedy, Goldstein 2003 — Studdart-Kennedy M., Goldstein L. Launching Language: The Gestural Origin of Discrete Infinity. *Language Evolution*. Christiansen M., Kirby S. (eds.). Oxford: Oxford Univ. Press, 2003, 235–254.

Wang et al. 2014 — Wang W., Arora R., Livescu K. Reconstruction of articulatory measurements with smoothed low-rank matrix completion. *Proc. of the 2014 IEEE Spoken Language Technology Workshop (South Lake Tahoe, Dec. 7–8, 2014)*. 54–59.

Wei, Dang 2008 — Wei J., Dang J. Vocal tract normalization in articulatory space using thin-plate spline method. *Journal of the Acoustical Society of America*, 2008, 123, 3885.

Westbury 1994 — Westbury J. R. *X-ray microbeam speech production database. User's handbook. Version 1*. Madison: Univ. of Wisconsin, 1994.

Yan et al. 2014 — Yan H., Dang J., Cao M., Kröger B. A new framework of neurocomputational model for speech production. *Proc. of the 9th International Symposium on Chinese Spoken Language Processing (Singapore, Sept. 12–14, 2014)*. 2014, 294–298.

Zhou et al. 2008 — Zhou X., Espy-Wilson C., Boyce S., Tiede M., Holland Ch., Choe A. A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /ɹ/. *Journal of the Acoustical Society of America*, 2008, 123(6): 4466–4481.

Получено / received 23.09.2023 *Принято / accepted 21.11.2023*