Информационные процессы, Том 5, №2, стр. 101 – 119.

====== ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ТЕХНИЧЕСКИХ ======
====== И СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ СИСТЕМАХ =======

Кодовая книга для речевых обратных задач

А.С.Леонов[*], И.С.Макаров[**], В.Н.Сорокин[**], А.И.Цыплихин[**]

[*]*Московский инженерно-физический институт, Москва, Россия*
[**]*Институт проблем передачи информации, Российская академия наук, Москва, Россия*
Поступила в редколлегию 19.03.2005

Аннотация - Обратная задача нахождения формы речевого тракта (или артикуляторных параметров) по акустическим данным сводится к поиску условного минимума некоторой целевой функции. В силу неоднозначности отображения пространства акустических параметров в пространство артикуляторных параметров такая задача минимизации является многоэкстремальной. Отбор наилучшего решения осуществляется в результате многократного запуска процесса оптимизации с начальными приближениями, выбранными специальным образом. Эти начальные приближения составляют кодовую книгу. Формирование кодовой книги само по себе требует решения некоторой обратной задачи. Ее решение, однако, облегчается возможностью использования траекторий некоторых точек внутри речевого тракта, измеренных с помощью микролучевого рентгеноскопа или электромагнитного артикулографа синхронно с записью речевого сигнала. Входные акустические параметры и структура кодовой книги зависят от типа речевого сегмента – гласного, назального, фрикативного или смычки. Квазистационарные сегменты описываются значениями артикуляторных параметров, содержащихся в каждой ячейке квантованных акустических параметров. Переходные процессы, характерные для взрывных согласных, описываются последовательностью акустических и артикуляторных параметров на некотором интервале времени.

1. Введение

Ниже рассматривается речевая обратная задача, состоящая в нахождении функции площади поперечного сечения речевого тракта (или его формы в среднесагиттальной плоскости) по измеренным параметрам речевого сигнала. Возможны варианты этой задачи, в которых по тем же данным ищутся артикуляторные параметры или управления артикуляцией. Математически эта задача формулируется как задача поиска условного экстремума некоторого критерия оптимальности (целевого функционала) при различного рода ограничениях. Речевая обратная задача является многоэкстремальной из-за неоднозначности отображения пространства акустических параметров в пространство артикуляций. Кроме того, точки локальных и глобальных экстремумов этой задачи могут быть неустойчивы по отношению к возмущениям данных задачи. Стандартный подход к решению подобных некорректных экстремальных задач состоит в: 1) применении регуляризующих алгоритмов оптимизации целевого функционала; 2) многократном применении этих алгоритмов с использованием различных начальных приближений из некоторого специального их множества; 3) в последующем сравнении результатов оптимизации и анализе полученных решений. Это позволяет выбрать решение с наиболее подходящими характеристиками.

Подробная математическая постановка речевой обратной задачи в различных вариантах, а также алгоритмы ее решения рассматривались нами в [1, 2, 3]. В данной статье даются лишь некоторые важные дополнения к постановке задачи и к деталям алгоритма ее решения (см. раздел 2). Главный акцент делается на проблемы, связанные с созданием массива указанных начальных приближений.

Начальные приближения можно задавать, используя различные стратегии. Один из подходов связан с созданием так называемой кодовой книги, в которой содержатся векторы артикуляторных параметров и соответствующие им векторы акустических параметров. Если имеется адекватная модель речеобразования, то формирование кодовой книги может происходить путем многократного решения прямой задачи, т. е. вычисления акустических характеристик по артикуляторным

параметрам, перебираемым с некоторым шагом. Метод кодовой книги для решения речевых обратных задач был впервые предложен в [4], и затем использовался в [5 - 9]. Авторы статьи [10] добились существенного уменьшения объема кодовой книги, используя только те артикуляторные векторы, которые находятся на траекториях переходных процессов от одного артикуляторного состояния к другому. Локальная линеаризация модели речеобразования для каждой ячейки кодовой книги позволяет существенно ускорить процесс решения обратной задачи [11].

В формировании кодовой книги значительную роль играют анатомические параметры речевого тракта, для которого эта книга составляется. Если кодовая книга построена для определенного диктора с известными размерами тракта, то обратная задача для этого диктора, например, при определении формы речевого тракта для гласных, может быть решена с высокой точностью. Однако попытка использования геометрических параметров одного диктора в решении обратной задачи для другого, скорее всего, приведет к неудаче [12]. Этот факт вызывает серьезные сомнения относительно эффективности использования кодовых книг, построенных с помощью абстрактных, не апробированных моделей, при решении реальных обратных задач для произвольного диктора.

Способ построения кодовой книги также влияет на эффективность решения обратной задачи. Во-первых, математическая модель артикуляции должна быть достаточно подробной с тем, чтобы обеспечить генерацию всех возможных в данном языке конфигураций речевого тракта. Во-вторых, вычисленные с помощью акустической модели параметры речевого сигнала должны соответствовать реальным параметрам. В идеале кодовая книга должна была бы строиться путем непосредственного измерения 3-хмерной формы речевого тракта в динамике синхронно с записью речевого сигнала. Однако пока такой измерительной техники не существует.

С появлением методов MRI стало возможным дальнейшее совершенствование моделей речеобразования, а установки типа микролучевых рентгеноскопов или артикулографов открыли пути к созданию кодовых книг, основанных на реальных артикуляторных и акустических данных. Синхронное измерение акустических параметров и координат нескольких точек на языке, губах, верхних и нижних зубах позволяет поставить задачу формирования кодовой книги для реальных дикторов как специфическую обратную задачу. Сложность этой задачи значительно меньше, чем для обратной задачи, входными данными которой служат только акустические параметры, особенно в том случае, когда кроме точек доступны и измерения формы твердого неба в среднесагиттальной плоскости, формы челюсти в латеральной плоскости, расстояния от передних зубов до задней поверхности тракта и положения гортани. Принципиальным вопросом является определение того, какие акустические параметры должны входить в кодовую книгу. До сих пор основное внимание в этом отношении уделялось формантным частотам, по которым искалась форма речевого тракта. Было установлено, что число формант при решении обратной задачи для гласных должно равняться трем [12]. Однако общая постановка обратной задачи дается не только для гласных, но и для назальных звуков, для взрывных на интервале смычки и для фрикативных, где форманты могут отсутствовать. В [13] для решения обратной задачи для фрикативных в качестве акустической информации использовался нормированный энергетический спектр. Вместе с тем, известно, что спектр подвержен влиянию АЧХ канала и аддитивных шумов. К тому же форма спектра фрикативных сильно меняется от диктора к диктору и в зависимости от контекста. Поэтому в работе [3] в результате специального исследования были найдены более устойчивые акустические параметры спектра фрикативных.

Выбор акустических параметров важен для успеха решения обратной задачи в реальных условиях. Поэтому создание кодовой книги (обучение) и решение обратной задачи для неизвестного речевого сигнала должны выполняться с использованием одних и тех же методов анализа и синтеза речевого сигнала. Те акустические параметры, которые в автоматическом режиме измеряются неустойчиво или с большой погрешностью, не могут быть использованы для формирования кодовой книги.

2. Модель артикуляции

Математическая модель артикуляции, которая использовалась для генерации формы речевого тракта, опирается на анализ действия внутриротовых и лицевых мышц. Она несколько отличается от модели, описанной в [14]. Основные мышцы, участвующие в артикуляции, и направление развиваемых ими усилий показаны на Рис. 1.

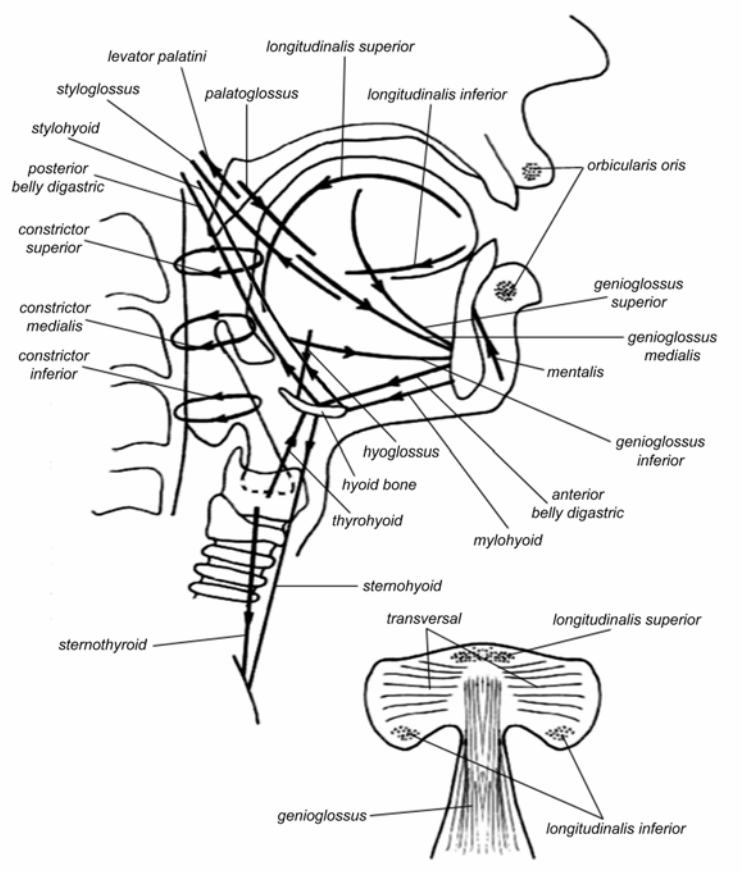

Рис.1. Схема мышц

Рассмотрим основные артикуляторы, входящие в модель.

Опускание небной занавески осуществляется за счет сокращения мышцы *palatoglossus*, а подъем - мышцей *levator palatini*. И хотя при своем подъеме небная занавеска деформируется как упругое тело, с точки зрения фонетических функций ее движения можно аппроксимировать лишь одним параметром - углом поворота небной занавески относительно поднятого положения, которое принимается за исходное. Этот угол определяет площадь прохода в носовую полость и, соответственно, степень назализации.

В процессе артикуляции высота гортани относительно твердого неба может меняться. Сокращение мышцы *sternothyroid* опускает гортань, а подъем гортани происходит при сокращении мышцы *thyrohyoid.* Три мышцы-сжиматели глотки *constrictor superior, constrictor medialis, constrictor inferior* определяют площадь ее поперечного сечения. В области между входом в пищевод и небной занавеской ширина глотки описывается коэффициентами при двух главных компонентах, полученных путем статистического анализа MRI данных.

Положение нижней челюсти описывается двумя параметрами - углом поворота относительно челюстного сустава и смещением челюсти вперед - назад. Угол поворота принимается равным нулю при сомкнутых зубах. Опускание нижней челюсти происходит при сокращении мышц *anterior belly digastric, posterior belly digastric, mylohyoid* при условии, что *hyoid bone* удерживается в фиксированном положении противодействующими мышцами. Подъем нижней челюсти создается сокращением мышц *temporalis* и *masseter*. Последние две мышцы совместно с *lateral pterygoid* также сдвигают нижнюю челюсть вперед, а при сокращении *posterior belly digastric* и задних волокон мышцы *temporalis* челюсть может сдвигаться назад. Мышцы *masseter, lateral pterygoid* и *temporalis* - внешние лицевые мышцы и на Рис. 1 не показаны.

Корень языка опускается при сокращении мышц *sternohyoid* и *hyoglossus*, а поднимается при сокращениии мышцы *stylohyoid*. Сокращение мышц *constrictor medialis* и *constrictor inferior* приближает корень языка к задней стенке речевого тракта, а сокращение *mylohyoid* и *genioglossus inferior* сдвигает его в сторону подбородка. Таким образом, положение корня языка описывается вертикальной и горизонтальной координатой.

Нижняя губа поднимается при сокращении мышцы *mentalis* и опускается при согласованном сокращении двух ветвей мышцы *depressor labii*. Верхняя губа поднимается при согласованном сокращении двух ветвей мышцы *levator labii*. Кроме того, вдоль обеих губ расположены волокна

кольцевой мышцы *orbicularis oris*, сокращение которой приводит к сближению углов губ и их выпячиванию, если мышца *buccinator* не активна. Мышцы *levator labii, depressor labii* и *buccinator* также относятся к внешним лицевым мышцам и на Рис. 1 не показаны. Вертикальное смещение верхней губы регистрируется далеко не у всех дикторов и не во всех контекстах, поэтому часто можно ограничиться только двумя параметрами - вертикальным положением нижней губы и горизонтальным смещением губ.

Мышца *styloglossus* охватывает нижнюю поверхность языка. Сокращение этой мышцы может привести к повороту языка как твердого тела относительно его корня.

Анатомически язык представляет собой изогнутую пластину, упругие деформации которой происходят под воздействием внутренних и внешних мышц. К этой пластине присоединены внешние мышцы, масса которых мало влияет на движения языка. Внешние мышцы языка - это *constrictor superior, palatoglossus, styloglossus, hyoglossus, genioglossus*. В мышце *genioglossus* различают три основных отдела: *superior, medialis, inferior*. Имеются также продольные внутренние мышцы *longitudinalis superior, longitudinalis inferior*, которые главным образом поднимают или опускают кончик языка, и поперечные мышцы *transversalis*. На Рис. 1 справа внизу показан поперечный разрез языка примерно в области *genioglossus medialis*. В поперечном сечении языка можно видеть мышцу *transversalis*, которая изгибает язык в поперечном направлении, создавая как выпуклость, так и впадину. Этот изгиб описывается "поперечной собственной функцией" в виде половины синусоиды. Этот параметр деформирует поверхность передней трети языка во фронтальной плоскости.

Упругие деформации языка под воздействием сокращения внешних и внутренних мышц описываются дифференциальным уравнением в частных производных в полярной системе координат, поскольку в нейтральном состоянии форма языка близка к полуокружности с радиусом R_0. Решение этого уравнения можно представить в виде произведения двух функций, одна из которых зависит только от угла в полярной системе координат, а другая - только от времени. Тогда поверхность языка описывается как

$$u(\varphi,t) = \sum_{k}^{5} a_k \psi_k(\varphi) T_k(t) + R_0,$$

где φ – угол отсчета поверхности языка в полярной системе координат, ψ_k - собственные функции упругих деформаций, T_k – временные моды. Подгонка решения уравнения упругих деформаций языка к измерениям поверхности языка на рентгенограммах показала, что для тела языка достаточно использовать 4 собственные функции [14]. Кончик языка часто выступает как самостоятельный артикуляторный орган, и для него достаточно использовать только одну собственную функцию. Таким образом, форма языка описывается пятью собственными функциями. Однако коэффициенты a_k при этих собственных функциях коррелированны, поскольку форма языка в основном управляется сокращением различных отделов мышцы *genioglossus*. Поэтому входными параметрами для первых четырех собственных функций упругих деформаций тела языка являются три параметра сокращения основных отделов мышцы *genioglossus,* и коэффициенты a_k определяются как

$$a_k(t) = \sum_{n=1}^{3} g_n(t) \int_{\varphi_{n-1}}^{\varphi_n} W_n(\varphi) \psi_k(\varphi) d\varphi ,$$

где g_n – усилие, развиваемое n-м отделом мышцы *genioglossus*, W_n – распределение усилий в n-м отделе, φ_0=0. Данные для пересчета мышечных усилий в коэффициенты при собственных функциях тела языка приведены в Табл. 1.

Табл. 1. Коэффициенты возбуждения собственных функций тела языка при сокращении волокон трех отделов мышцы *genioglossus*.

собственные функции языка	1	2	3	4
собственные частоты (рад/сек)	28	52	82	10
нижний отдел *genioglossus*	0.41	-0.06	-1.0	0.44
средний отдел *genioglossus*	-0.1	0.84	-0.2	-1.0
верхний отдел *genioglossus*	-0.7	0.61	0.94	0.59

На начальном этапе создания кодовой книги приходится решать обратные речевые задачи, практически не имея надежных данных о величинах искомых артикуляторных параметров. Это

справедливо даже в случае, если имеется информация о координатах некоторых точек внутри речевого тракта. Поэтому для этих артикуляторных параметров необходимы некоторые начальные приближения, с которых может стартовать процесс минимизации критерия оптимальности, заложенный в алгоритм решения обратной задачи. Такие начальные приближения удобно находить интерактивно с использованием визуализации артикуляторной модели (Рис. 2). Манипулируя артикуляторными параметрами, можно создать такую форму речевого тракта, в которой рассогласование между измеренными координатами опорных точек и движущимися поверхностями тракта будет невелико.

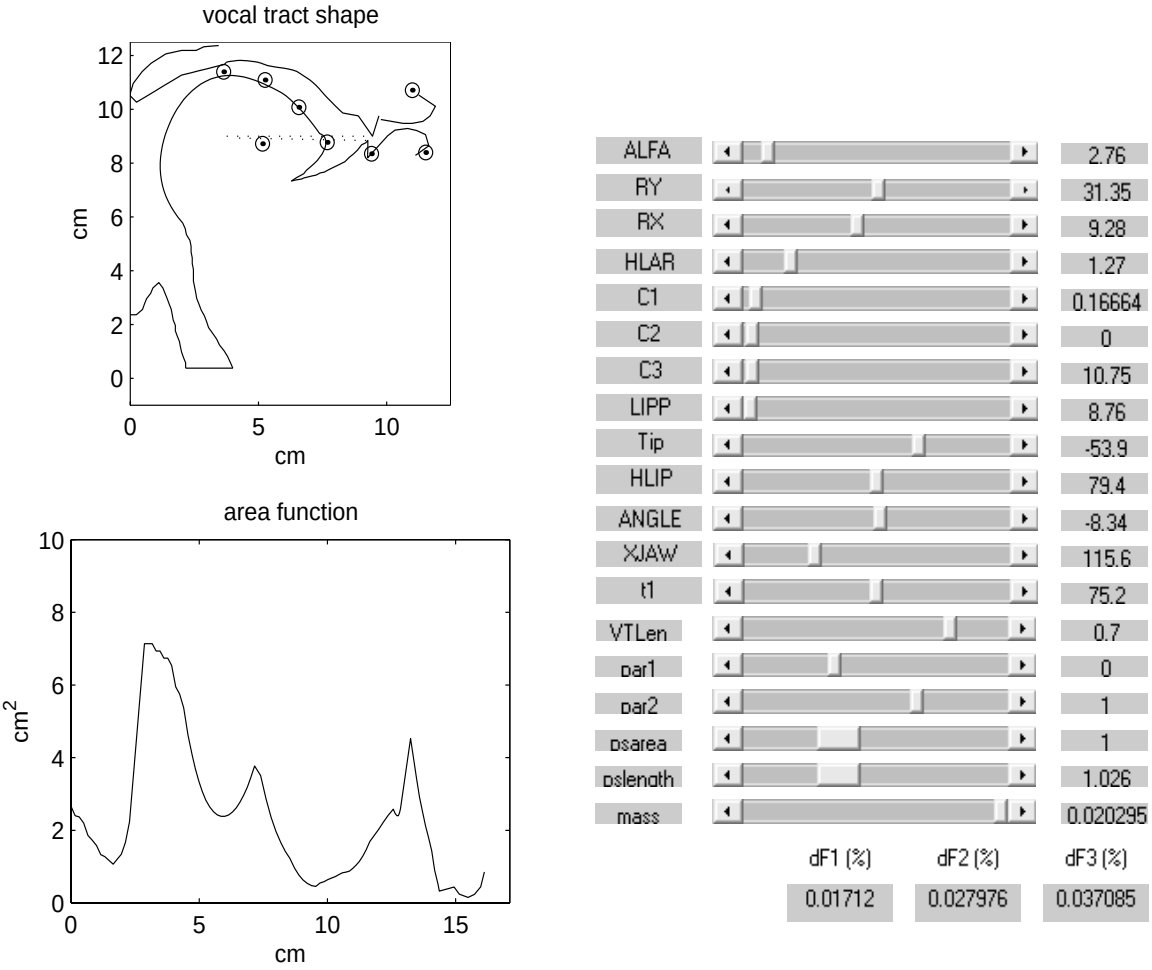

Рис. 2. Графический интерфейс программы для подбора начальных приближений.

При решении обратных речевых задач вариационным методом вычисляется невязка между измеренными и вычисленными параметрами речевого сигнала. При этом не только определяются резонансные частоты речевого тракта, но в том или ином виде выполняется синтез речевого сигнала. Для фрикативных и взрывных согласных необходимо использование специальной модели аэродинамических процессов ("по постоянному току", в отличие от акустических колебаний). Методы вычисления резонансных частот, в том числе и для разветвляющегося речевого тракта, описаны в [14 - 16]. Аэродинамические процессы описаны в [15].

3. Сегментация речевого сигнала

Ранее было установлено, что решение обратной задачи для разных типов звуков требует использования разных критериев оптимальности. К тому же при решении обратной задачи, как и в процессе создания кодовой книги, необходимы некоторые артикуляторные параметры, отсутствующие в базе данных, составленной по результатам измерения артикуляторных движений синхронно с речевым сигналом. К числу этих параметров относится высота небной занавески, отличающая назальные смычки и назализованные гласные от звонких смычек и ротовых гласных. Кроме того, необходима информация о наличии голосового возбуждения и о ширине голосовой щели

для глухих фрикативных и взрывных согласных. Эти параметры определяются путем акустического анализа речевого сигнала. В результате этой процедуры сигнал сегментируется на участки, размеченные принадлежностью к одному из четырех типов: гласноподобные, назальные, фрикативные и смычки, причем отдельно указывается наличие или отсутствие голосового возбуждения. Процедура сегментации по существу представляет собой процесс распознавания типа сегмента. В определенной степени эта задача проще, чем распознавание фонетических элементов. Однако и здесь приходится использовать сложные правила принятия решений вследствие разнообразия акустических характеристик сегментов (даже одного и того же типа).

От точности определения типа сегмента и его временных границ существенно зависит и качество решения обратной задачи, использующей эти данные. Построение процедуры сегментации весьма сложно. Оно включает в себя обучение с помощью специальной базы данных, заранее размеченной вручную на основные фонетические элементы изучаемого языка. При этом для каждого типа сегмента необходимо найти информативные параметры. В пространстве этих параметров выполняется аппроксимация их распределения вероятности суперпозицией нормальных распределений и формируется алгоритм принятия решений. Детальное описание процесса сегментации требует отдельного рассмотрения, и в данной работе приводятся только примеры работы сегментатора (Рис. 3).

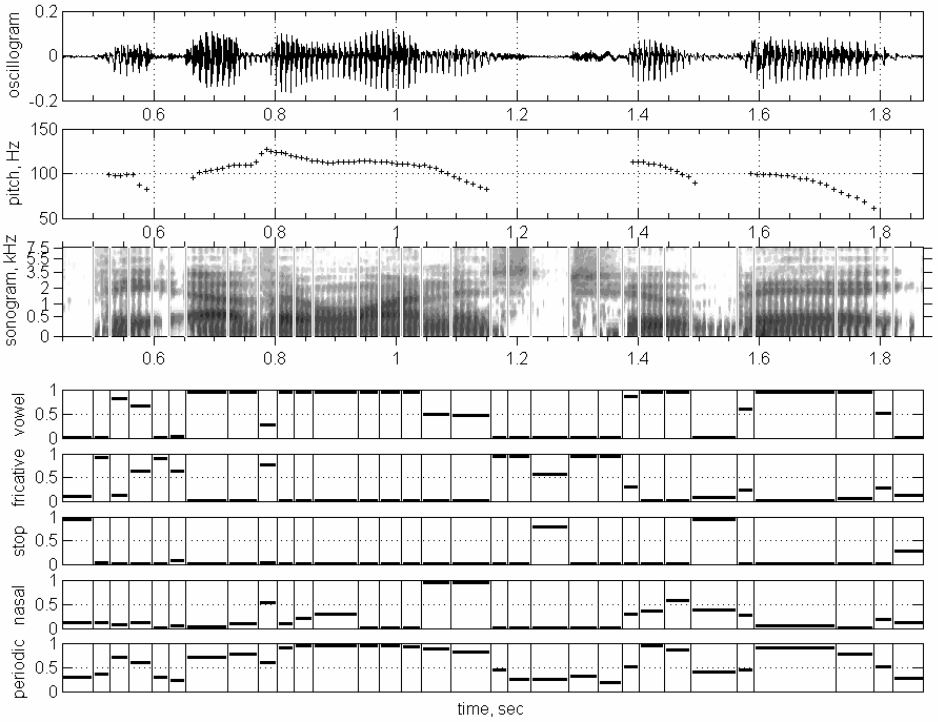

Рис. 3. Сегментация фразы "The other one is too big" с указанием апостериорной вероятности присутствия сегмента каждого типа.

Необходимо, однако, отметить, что сегментация речевого сигнала на основные типы в автоматическом режиме не может быть выполнена со 100% надежностью. Наиболее часто встречающиеся типы ошибок – это решение о наличии гласного на сегменте назальной смычки, и наоборот. Поскольку гласные и назальные звуки описываются формантными частотами, то ошибки в типе сегмента могут привести к фатальным ошибкам в решении обратной задачи (типа неправдоподобной формы речевого тракта). Отсюда следует необходимость решения обратной задачи в предположении, что данный сегмент речевого сигнала может принадлежать любому из четырех типов. Выбор окончательного решения из полученных четырех должен выполняться на основе сравнения полученных невязок заданных и вычисленных акустических параметров, а также по усилиям, необходимым для формирования найденного вектора артикуляторных параметров при переходе от предыдущего сегмента.

4. Структура кодовой книги

Помимо обеспечения начальными приближениями для решения обратной задачи, кодовая книга может использоваться в ряде других задач (например, при артикуляторном синтезе и распознавании речи). Поэтому, кроме акустических и артикуляторных параметров, в ней должна содержаться информация о фонетической принадлежности артикуляторных и акустических векторов. Следует также различать содержание кодовой книги для стационарных состояний артикуляции и переходных процессов, а также процесс формирования кодовой книги и ее использование.

При входе в кодовую книгу со стороны акустических параметров для стационарных состояний необходимо квантовать акустические параметры таким образом, чтобы образовалось сравнительно небольшое число ячеек, которые содержат артикуляторные векторы, соответствующие данному акустическому вектору. Шаг квантования определяет число ячеек и объем артикуляторных векторов, содержащихся в каждой ячейке. Очевидно, что для формантных частот минимальное значение шага квантования равно порогу восприятие по частоте. Согласно [17], порог восприятия формантных частот близок к Δ_f= 3 - 5%. В [18] приводятся оценки этого порога для разных формант: 10 – 30 Гц для первой форманты и 20 – 100 Гц для второй форманты, что дает примерно такие величины. Однако погрешность автоматического определения формантных частот в технических системах значительно превышает порог восприятия, достигая, по некоторым оценкам, 10% и даже выше [19, 20]. Тестирование нашего алгоритма анализа формантных частот было проведено путем обработки сигнала, синтезированного для известных треков формант. Была получена оценка погрешности в 27 Гц (среднее значение 5.9%) для F_1, 37 Гц (среднее значение 2.6%) для F_2, 28 Гц (среднее значение 1.2%) для F_3. Определить погрешность анализа характерных частот спектра фрикативных звуков достаточно трудно в силу значительного разнообразия их характеристик. Поэтому для них был принят шаг квантования, несколько превышающий Δ_f, а именно Δ_s = 8%.

В [10] было показано, что запуск процесса оптимизации из разных начальных приближений значительно увеличивает вероятность нахождения если не глобального, то локального экстремума, обеспечивающего приемлемую ошибку как по входным акустическим, так и по артикуляторным параметрам. Артикуляторные параметры, попадающие в одну и ту же ячейку, должны быть квантованы. Это нужно не только для сокращения объема кодовой книги, но и для того, чтобы начальные приближения для процесса оптимизации заметно отличались. Шаг квантования можно примерно оценить из экспериментов по ресинтезу речи с разной точностью представления артикуляторных параметров. В работе [1] было найдено, что квантование артикуляторных треков на 256 уровней практически не ухудшает аппроксимацию по сравнению с представлением артикуляторных параметров в формате с плавающей запятой. Это означает, что для синтеза речи артикуляторные параметры должны быть представлены с точностью около 0.4%, что в свою очередь указывает на примерную точность представления артикуляторных параметров в кодовой книге.

4.1. Статическая кодовая книга

Статическая часть кодовой книги основана на анализе акустических и артикуляторных параметров на сегментах речевого сигнала, на которых спектральный состав звука мало меняется. К таким сегментам относятся звонкая, глухая и назальная смычка, а также участки фрикативных и гласных звуков. Метод разбиения речевого сигнала на квазистационарные сегменты был описан в статье [21]. Во всех известных публикациях, посвященных формированию кодовой книги, процесс ее создания направлен от артикуляции к акустике, т.е. акустические параметры вычисляются как результат вариации параметров модели речеобразования. Это позволяет избежать трудностей в измерении акустических параметров на этапе формирования кодовой книги, но допускает существование таких сочетаний артикуляторных и акустических параметров, которые не встречаются в реальности. Поэтому и формирование кодовой книги, и ее использование должны выполняться с учетом особенностей анализа акустических параметров. При этом проблема заключается не только в точности определения акустических параметров, но и в идентификации самих этих параметров.

В одном и том же частотном диапазоне речевого сигнала может наблюдаться разное число резонансов. Дополнительные резонансы появляются при опускании небной занавески, когда речевой тракт разветвляется в носовую полость. Резонансные колебания в подсвязочной области – трахее, бронхах и легких также могут либо проникать в ротовую полость при раскрытой голосовой щели, либо регистрироваться микрофоном вследствие излучения через ткани шеи и грудной клетки. Наблюдаются также случаи пропадания оценок формантных частот в какие-то моменты времени. Естественно, что резонансная частота речевого тракта при этом никуда не исчезает, а просто ухудшаются условия ее определения конкретным алгоритмом. На Рис. 4 показан пример появления "лишней" форманты на интервале времени 4.55 – 4.65 с.

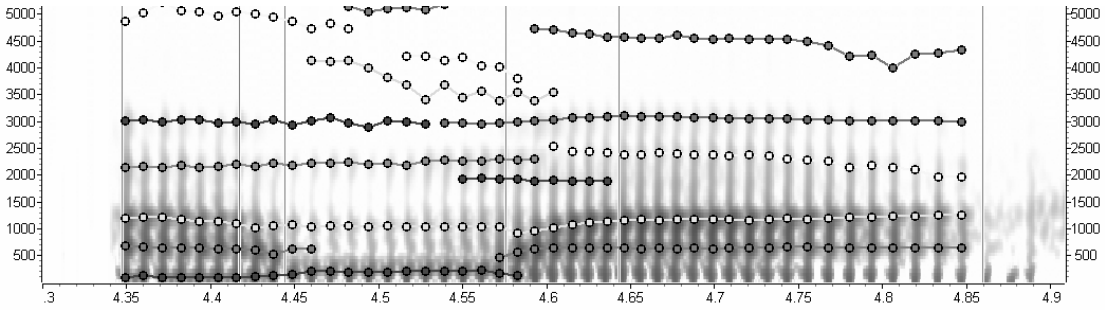

Рис. 4. Сонограмма и треки формант для слога /*ама*/.

Пример пропадания оценок третьей форманты на переходных процессах в окрестности смычки показан на Рис. 5.

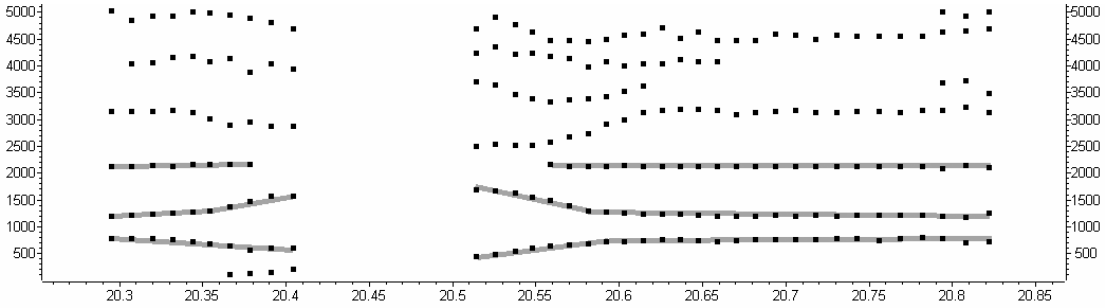

Рис. 5. Треки формант для слога /*ада*/.

Имеется целый ряд причин, вызывающих подобные погрешности в определении формантных частот. Амплитуда резонансных колебаний может упасть до нуля при попадании импульса голосового возбуждения в противофазу этих колебаний. К аналогичному эффекту могут привести биения в сигнале вследствие интерференции волн в помещении с реверберацией. Если амплитудно-частотная характеристика микрофона имеет нули или области с малым коэффициентом усиления, то при попадании резонансной частоты речевого тракта в эти области измеренная энергия форманты уменьшается или исчезает совсем. Быстрое изменение резонансной частоты при артикуляции согласных звуков приводит к увеличению ее эффективной полосы и к понижению энергии на частоте резонанса. Рассинхронизация момента определения формантной частоты относительно интервала закрытой голосовой щели приводит как к изменению этой частоты, так и к падению ее энергии. Кроме того, при определенных условиях резонансные частоты сближаются настолько, что становятся неразличимыми (Рис. 6).

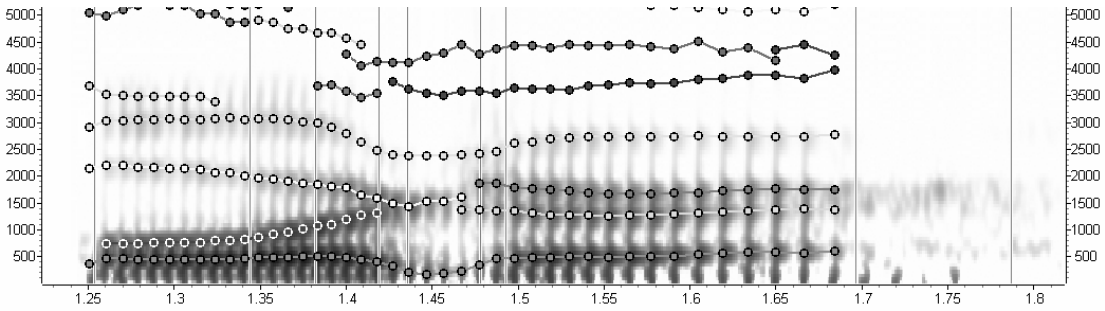

Рис. 6. "Слияние" второй и третьей формант.

Поскольку при вычислении невязки между вычисленными и измеренными резонансными частотами важна нумерация резонансов, то неустойчивость в определении порядкового номера форманты

создает серьезные трудности как при формировании кодовой книги, так и при обращении к ней в процессе решения обратной задачи. Отсюда следует, что мгновенные оценки формантных частот, даже и синхронно с основным тоном, не применимы как к формированию, так и к использованию кодовой книги.

Один из способов компенсации ошибок анализа формант состоит в усреднении данных на некотором интервале, на котором спектральный состав звука меняется мало. Усреднение данных выполняется путем вычисления среднего профиля спектра и последующего анализа формантных частот автокорреляционным методом линейного предсказания. Это позволяет отсеять некоторые случайные появления лишних оценок формант или возместить их отсутствие. Если же появление или исчезновение формант связано не с ошибками анализа, а с реальными акустическими процессами, то необходимо решать обратную задачу для разных вариантов сортировки формант с последующим выбором наиболее подходящего решения. В частности, основанием для такого выбора может служить распределение амплитуд формант. Невязка между координатами опорных точек и вычисленных точек на поверхности речевого тракта также дает основание для выбора между различными решениями обратной задачи на сегменте с неоднозначной сортировкой формант.

Один из способов избежать трудности в определении формантных частот состоит в использовании кепстральных коэффициентов для описания формы спектрального профиля [9]. Напомним, что кепстром называется обратное преобразование Фурье от логарифмированного энергетического спектра, обычно представленного в шкале частот *мел*. При этом также в значительной степени ослабляется проблема сегментации речевого сигнала на основные типы. Однако при всем удобстве такого формального подхода теряется физическая интерпретируемость кепстральных коэффициентов и существенно снижается способность оптимизатора к дифференцированному подходу к физически разным акустическим процессам. В частности, кепстральные коэффициенты дают интегральное описание формы спектра и не позволяют определить форму спектра в определенных частотных областях.

Поэтому, в дополнение к акустическим данным в виде формантных частот, целесообразно использовать и какую-то форму описания спектрального профиля, усредненного на интервале стационарности. При этом, однако, примерно на два порядка увеличивается объем входных данных, поскольку спектр содержит более 100 отсчетов, и существенно усложняется и замедляется процесс выбора артикуляторных векторов. Кепстральное описание не намного более экономно, т.к. обычно используется до 40 кепстральных коэффициентов.
В каждой акустической ячейке для гласных записываются и амплитуды формант, а также средний профиль спектра, определенный на сегменте стационарности.

Нередко наблюдаются случаи, когда участок речи разбивается на несколько квазистационарных сегментов, хотя фактически все эти сегменты принадлежат одному и тому же типу, например, гласному или фрикативному. Тогда появляется еще одна возможность корректировки входных данных акустического анализа путем сравнения решений обратной задачи на соседних сегментах. Вычисляя силы, необходимые для перехода из одного артикуляторного состояния в другое, и пользуясь критерием минимума этих сил, а также ограничениями на максимальные мышечные усилия, среди всевозможных решений на соседних сегментах отбираются только такие, которые удовлетворяют ограничениям и критерию оптимальности. О вычислении управляющих сил говорится в разделе 5. Дополнительный контроль правильности решения обратной задачи обеспечивается ресинтезом речевого сигнала и сравнением его с оригинальным звукосочетанием.

Проблема множественных решений возникает также при создании кодовой книги не для одного, а для группы дикторов. Можно ожидать, что в одной и той же ячейке акустических параметров окажутся разные артикуляторные вектора для разных анатомических параметров. Тогда, например, в задачах распознавания или сжатия речи необходимо, прежде чем принять решение о форме речевого тракта или командах управления артикуляцией, определить тип диктора, не допуская перескока с одного типа на другой для отрезка слитной речи. Это тем более актуально в задачах верификации или идентификации диктора на артикуляторном уровне.

В зависимости от предполагаемого типа сегмента используются разные акустические параметры. Как уже упоминалось, обратная задача для гласных требует знания трех формантных частот. Определяя шаг по частоте формант как максимум между 32 Гц и 5% от частоты, получим, что максимальное количество ячеек в кодовой книге, равное 19*40*22=16720. Эта оценка заведомо завышена, поскольку не учитывает взаимной зависимости формант. Кодировка формантных частот используется для входа в кодовую книгу. Но в каждой ячейке записываются также средние амплитуды формант и средний спектр. Эти данные помогают процессу оптимизации. В частности, при неправильной сортировке формант ошибка может быть обнаружена по рассогласованию с амплитудами.

Наблюдается существенное различие в спектрах фрикативных разного типа. Поэтому и число входных акустических параметров для фрикативных оказывается больше, чем для гласных. Опишем эти параметры. Для спектрального профиля фрикативного, усредненного на интервале стационарности, вычисляется линейная функция, аппроксимирующая спектр. С помощью этой линии находятся 5 характерных частот: три частоты пересечения средним профилем спектра этой линии при опросе от высоких частот и при опросе от низких частот, а также частота центра тяжести в области высоких частот, амплитуда которых выше значений линейной функции. Высокие фрикативные описываются двумерным пространством, которое содержит примерно 200 ячеек, а низкие фрикативные распознаются в другом, трехмерном, пространстве, содержащем около 350 ячеек. Аналогично амплитудам формант для гласных в каждой ячейке записывается и коэффициент наклона этой линейной функции, а также средний спектр.

На сегменте назальной смычки может наблюдаться от одной до трех формант в диапазоне частот до 3000 Гц. Поэтому вход в кодовую книгу, аналогично гласным, кодируется по трем формантам, однако допускается присутствие лишь одной или двух формант. Амплитуды формант также содержатся в каждой ячейке, но уже как дополнительная информация. Если сегмент определен как назальный, то в каждой ячейке высота небной занавески задается таким образом, чтобы проход в носовую полость составлял не менее 0.05 $см^2$.

Смычка характеризуется либо полным отсутствием энергии на всех частотах (для глухой смычки), либо энергией в низкочастотной области (для звонкой смычки). Соответственно, акустическим входом в слой звонкой смычки является частота так называемого радиального резонанса в области низких частот. Помимо этого, используются параметры взрыва, о которых говорится ниже в разделе 4.2.

В каждом слое статической кодовой книги, для каждой ячейки акустических параметров, помимо множества векторов артикуляторных параметров, записываются координаты участка с минимальной площадью сечения тракта и значение этой площади, а также число Рейнольдса. В частности, для взрывных и назальных указывается координата участка речевого тракта с нулевой площадью поперечного сечения. Кроме того, каждой ячейке приписывается символ фонетической транскрипции квазистационарного сегмента.

4.2. Динамическая кодовая книга

Процесс речеобразования носит динамический характер, и доля стационарных сегментов в слитной речи относительно невелика. Сравнительно медленные переходные процессы, например, между гласными, еще могут рассматриваться как последовательность квази-стационарных состояний, и подобный подход применим к решению обратной задачи (примеры подобных решений приводятся в [9, 22]). В общем же случае описание переходных процессов с помощью мгновенных оценок акустических параметров, преимущественно формантных частот, нецелесообразно в силу рассмотренных выше явлений неустойчивости оценки этих частот. На переходном участке нельзя воспользоваться усреднением акустических параметров для повышения устойчивости анализа. Сегмент, стационарный с точки зрения алгоритма, оценивающего однородность спектра, вполне может содержать формантные переходы. Переходный процесс может отмечаться наличием формантных переходов или изменением общей энергии, и начало или конец команды на управление артикуляцией должен определяться по меткам начала и конца динамического сегмента. Первые попытки решения динамической обратной задачи для взрывных описаны в работе [9]. Было показано, что квазистационарный подход к решению динамической обратной задачи в общем случае приводит к неудовлетворительным результатам.

Некоторая помощь в регуляризации формантных треков на переходных процессах может быть получена с помощью линейной аппроксимации трека в смысле минимума среднеквадратичной ошибки. Пример такой аппроксимации показан на Рис. 5. К тому же для одного и того акустического вектора, взятого на траектории перехода из одного состояния в другое, могут встречаться артикуляторные вектора, различающие по величине и скорости изменения компонент. Таким образом, сегменты с переходными процессами должны описываться в терминах динамики, а не статики.

Формирование кодовой книги для динамических процессов значительно сложнее, чем для статики. Если некоторый сегмент определен как смычка, неважно, глухая, звонкая или назальная, то это указывает лишь на существование участка в речевом тракте с нулевой площадью поперечного сечения. Место же положения этого участка, так называемое "место артикуляции", определить невозможно, поскольку на глухой смычке вообще никакой сигнал не излучается, звонкая смычка лишь указывает на наличие голосового возбуждения (да и то не всегда), а назальная смычка, в основном, указывает на опущенную небную занавеску. Место артикуляции согласных звуков

проявляется в направлении движения второй и третьей формант в окрестности смычки, а также в спектральных характеристиках взрыва.

Роли формантных переходов и характеристик взрыва взаимодополняющие, причем для разных типов звуков может доминировать только один вид этой информации. Для звонких взрывных типа /б, д, г/ возбуждение голосового источника обычно не прекращается на интервале смычки, а взрыв довольно слабый и легко маскируется шумами канала связи. Поэтому в зависимости от типа гласного формантные переходы могут более или менее четко проявляться в речевом сигнале. Аналогично, для назальной смычки формантные переходы также могут наблюдаться, хотя их вид меняется при опускании небной занавески на сегменте гласного. Глухие взрывные типа /п, т, к/ образуются с раскрытием голосовой щели, часто еще до возникновения смычки, а после взрыва смычки проходит некоторое время до начала голосового возбуждения, так что формантные переходы могут не наблюдаться. В этом случае информация о месте артикуляции получается путем сравнения спектральных характеристик взрыва глухой смычки со спектром последующего гласного [23 - 28].

При анализе спектральных характеристик взрыва необходимо принимать во внимание влияние места артикуляции на амплитуды наблюдаемых резонансов. Как было показано в [14], возбуждение акустических колебаний в любом месте речевого тракта, кроме голосовой щели, сопровождается появлением нулей в спектре речевого сигнала. Это является следствием выражения для расчета амплитуды A_k возбуждения k-го резонансного колебания:

$$A_k(t) = \frac{2\int_0^l F(x,t)S(x,t)\psi_k(x)dx}{\rho_0\int_0^l S(x,t)\psi_k^2(x)dx},$$

где x – координата вдоль оси речевого тракта, l – длина речевого тракта, $F(x,t)$ – источник возбуждения, $S(x,t)$ – площадь поперечного сечения тракта, ψ_k – k-я собственная функция, ρ_0 – плотность воздуха. Импульсный источник возбуждения сосредоточен в узкой области пространства, поэтому его можно рассматривать как точечный, т.е. $F(x,t)=F(x_0,t)$, и

$$A_k(t) = \frac{2F(x_0,t)S(x_0,t)\psi_k(x_0)}{\rho_0}$$

Поэтому амплитуда возбуждения A_k зависит от места источника возбуждения, и если x_0 совпадает с одним из нулей собственной функции, то и соответствующие резонансные колебания не возбуждаются.

Необходимо также учитывать разницу в граничных условиях со стороны голосовой щели при артикуляции звонких и глухих согласных. Площадь голосовой щели для звонких согласных достаточно мала, так что в первом приближении граничные условия соответствуют жесткой стенке. При артикуляции глухих взрывных в момент взрыва голосовая щель раскрыта таким образом, что на акустические характеристики речевого тракта влияет подсвязочная область – трахея, бронхи и легкие. Поэтому и характеристики звонкого и глухого взрыва могут отличаться при одном и том же месте артикуляции.

Разница в наблюдаемости формантных переходов в зависимости от звонкости/глухости смычки иллюстрируется сонограммами слогов /ага/ и /ака/ (Рис. 7). Как видно, для глухого /к/ формантные переходы в окрестности смычки практически отсутствуют. Это является следствием установленного в [14] свойства почти закрытых труб, согласно которому скорость изменения резонансных частот при малой площади в области сужения значительно (на порядок) выше скорости изменения этой площади. Поэтому при определенной задержке между моментом размыкания смычки и возбуждением резонансных колебаний переходные процессы резонансных частот уже завершены и не несут информации о месте артикуляции.

В результате, в отличие от гласных и, отчасти, от фрикативных, информация о форме речевого тракта для взрывных согласных и назальных определяется не мгновенным (или средним на интервале стационарности) спектром, а распределена по окрестностям смычки, включая характеристики взрыва. Соответственно, и организация кодовой книги для решения обратной задачи для таких звуков гораздо сложнее, чем для гласных и фрикативных.

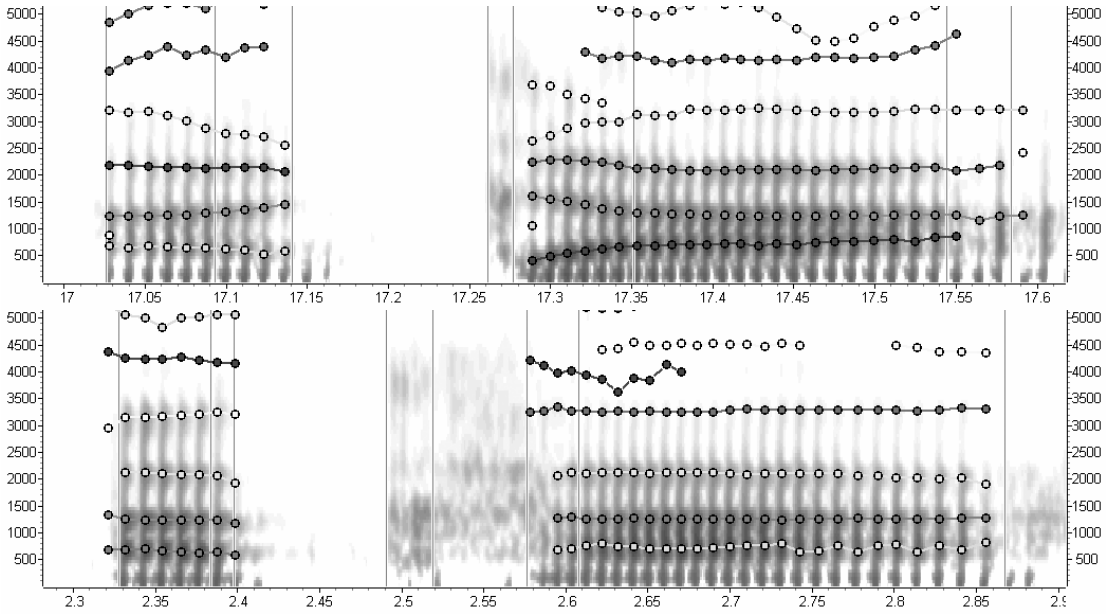

Рис. 7. Сонограммы и треки формант слогов /*ага*/ (вверху) и /*ака*/ (внизу).

Структура динамического слоя кодовой книги поясняется следующими качественными, но достаточно строгими соображениями. Известно, что при закрывании акустической трубы на одном конце, в то время как она уже закрыта на другом конце, частота первого резонанса стремится к нулю или к частоте радиального резонанса в трубе с податливыми стенками. Поэтому направление движения первой форманты практически не несет информации о месте смычки в речевом тракте. Частоты второго и третьего резонанса стремятся к разным величинам, в зависимости от формы трубы и от места ее перекрытия. Известно, что частота некоторого резонанса повышается или понижается в зависимости от того, приходится ли координата сужения на пучность или узел соответствующей собственной функции давления. Таким образом, сопоставляя резонансные частоты невозмущенной трубы и резонансные частоты непосредственно перед перекрытием трубы, можно определить место перекрытия. Согласно некоторым экспериментам по синтезу речи, форма трека (линейная, выпуклая или вогнутая) резонансной частоты вблизи смычки практически не влияет на восприятие места артикуляции согласного. Установлено также, что основное перцептивное влияние определяется не скоростью переходного процесса по частоте, а разностью частот в начале и конце переходного процесса. Поэтому и для создания кодовой книги целесообразно использовать оценки начальных и конечных значений формантных частот на сегменте. При этом линейная или параболическая интерполяция может помочь в определении целевых состояний.

Входными данными для динамической обратной задачи должен быть не один вектор акустических параметров (частот трех формант), а по крайней мере два вектора. Если гласный звук предшествует смычке, то первый вектор содержит значения формантных частот на интервале гласного перед началом формантного перехода или началом падения энергии речевого сигнала, а второй вектор содержит последние значения формантных частот непосредственно перед смычкой. Если смычка предшествует гласному, то это должны быть формантные частоты сразу после взрыва смычки и их значения на сегменте гласного после завершения переходного процесса. Если смычка завершается взрывом, то первый вектор акустических параметров должен содержать оценки частот второго и третьего резонанса, а второй – установившиеся значения формантных частот последующего гласного. Интервал времени между отсчетами этих векторов также важен для решения обратной задачи.

Во многих случаях взрывы смычек, особенно глухих, сопровождаются более или менее длительным участком так называемой аспирации, на котором действует турбулентный источник возбуждения, аналогичный источнику фрикативных. При этом характеристики аспиративного участка взрыва согласных /д, т/ близки к характеристикам фрикативных /с, ш/, характеристики аспиративного участка взрыва согласных /г, к/ близки к характеристикам фрикативного /х/, а характеристики аспиративного участка взрыва согласных /б, п/ близки к характеристикам фрикативного /ф/. Таким образом, для решения обратной задачи относительно места артикуляции взрывных и назальных в качестве входных акустических данных должны использоваться частоты F_1, F_2 и F_3 трех формант стационарного сегмента гласного перед смычкой и после смычки, F_2 и F_3 формантных переходов в

последний момент перед смычкой, интервал времени T_1 между отсчетами этих формант перед смычкой и после смычки, F_2 и F_3 взрыва или первого отсчета формант после смычки, интервал времени T_2 между отсчетами этих формант после смычки, а также характерные частоты аспиративного участка после взрыва (если таковой наблюдается). Дополнительная информация о звонкости/глухости смычки заключается в интервале времени T_3 между взрывом и началом фонации последующего гласного. Этот признак особенно важен для некоторых языков, в частности, для английского.

На акустические характеристики согласных влияет явление так называемой коартикуляции. Взрывные и назальные звуки не существуют отдельно. Обычно они нуждаются хотя бы в одном гласном, на фоне которого образуется смычка в речевом тракте. В слитной речи согласные звуки образуются на фоне перехода от одного гласного к другому, поэтому акустические характеристики перед смычкой и после нее зависят от предшествующего и последующего гласного. Наряду с этим, акустические характеристики согласного зависят и от индивидуальной тактики артикуляции, присущей конкретному диктору. В одних случаях переходный процесс от гласного к гласному совершается на интервале смычки и не влияет на акустические характеристики речевого сигнала в окрестности смычки. В других же случаях переходный процесс гласных начинается раньше или одновременно с движением к образованию смычки. Это явление и называется коартикуляцией. На этот процесс влияет также контекст и скорость артикуляции.

Отсюда следует, что кодовая книга для взрывных и назальных должна создаваться, по крайней мере, для пар типа "согласный-гласный" и "гласный-согласный", а в идеале и для всех возможных сочетаний "гласный-согласный-гласный". В этом смысле подготовка к решению обратной задачи для произвольного звукосочетания похожа на обучение системы автоматического распознавания речи.

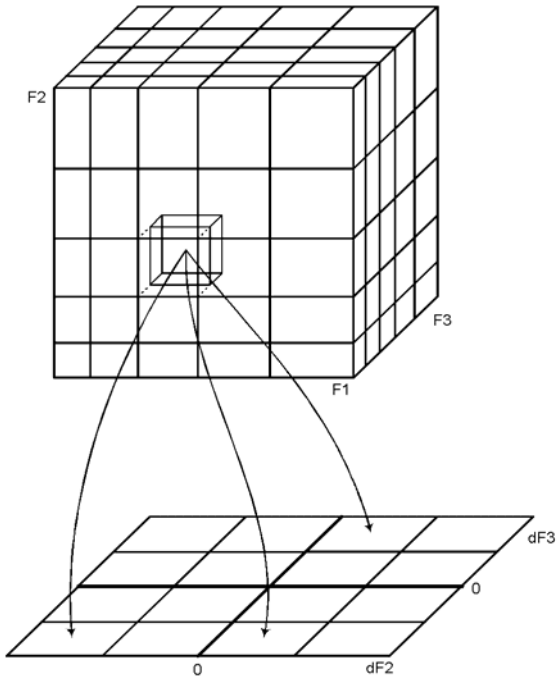

Рис. 8. Ячейка кодовой книги для стационарного сегмента гласного (вверху) и приращения формант для разных мест артикуляции.

Если смычке предшествует гласный, то входными данными в кодовую книгу являются формантные частоты F_1, F_2, F_3, либо усредненные на интервале стационарности гласного, либо измеренные в момент времени t_1 начала переходного процесса к смычке и формантные частоты F_2, F_3, измеренные в момент времени t_2 начала смычки. Начало переходного процесса определяется по скорости изменения любой из первых трех формант и скорости изменения общей энергии речевого сигнала. Если смычка звонкая, то может быть добавлено среднее значение радиального резонанса F_r, которое помогает разделить заднеязычную смычку от переднезычной и губной. Из каждой ячейки

квантованного по частоте вектора (F_1, F_2, F_3, t_1) происходит обращение к двумерному массиву с квантованными координатами $\Delta F_2 = F_2(t_2) - F_2(t_1)$, $\Delta F_3 = F_3(t_2) - F_3(t_1)$. В каждой ячейке этого массива находится множество артикуляторных векторов V_c^-, соответствующих форме речевого тракта непосредственно перед смычкой. Рис. 8 схематически иллюстрирует связь между формантными частотами на стационарном участке гласного и приращениями второй и третьей формант в зависимости от места артикуляции.

Каждый вектор V_c^- связан с вектором V_c в начале смычки в момент времени t_3 (обычно t_3 близко к t_2), а также интервалами времени $\Delta t_{12} = t_2 - t_1$, $\Delta t_{23} = t_3 - t_2$. По этим данным, а также по артикуляторному вектору, соответствующему началу переходного процесса в момент времени t_1, вычисляется команда (или последовательность команд), обеспечивающая приход на смычку в момент времени t_3. Кроме того, с каждой парой векторов связана координата x_0^-, в которой площадь поперечного сечения речевого тракта впервые приобретает нулевое значение в момент образования смычки.

В силу слабой связи между акустическими параметрами перед смычкой и местом артикуляции, во втором массиве могут быть представлены артикуляторные вектора и команды управления, соответствующие приходу на разные места артикуляции. Эта неоднозначность может быть разрешена путем анализа акустических параметров после взрыва смычки.

Если гласный следует за смычкой, то определяются резонансные частоты (F_2, F_3, t_1) в речевом сигнале в измеренные момент взрыва смычки t_1, резонансные частоты (F_1, F_2, F_3, t_2) начала переходного переходного процесса гласного, а также резонансные частоты (F_1, F_2, F_3, t_3) в момент окончания переходного процесса t_3. Аналогично обработке сигнала перед смычкой, происходит обращение к двумерному массиву с квантованными координатами $\Delta F_2 = F_2(t_3) - F_2(t_1)$, $\Delta F_3 = F_3(t_3) - F_3(t_1)$, в котором находятся пары артикуляторных векторов (V_c, V_c^+), соответствующие артикуляторному состоянию в последний момент смычки и в момент взрыва, а также интервалы времени $\Delta t_{12} = t_2 - t_1$, $\Delta t_{23} = t_3 - t_2$. По этим данным, а также по артикуляторному вектору V_w, соответствующему концу переходного процесса в момент времени t_3, вычисляется команда (или последовательность команд), обеспечивающая приход на квазистационарное состояние гласного в момент времени t_3. Кроме того, с каждой парой векторов связана координата x_0, в которой площадь поперечного сечения речевого тракта впервые приобретает нулевое значение непосредственно перед взрывом смычки. Во всех случаях, когда можно определить длительность смычки Δt_c от ее начала до взрыва, она также записывается в ячейку (ΔF_2, ΔF_3).

Смысл запоминания координат x_0^- и x_0^+ состоит в том, что они определяют место артикуляции, и это позволяет отсеять неправдоподобные переходы, например, в слогах "гласный1 – согласный – гласный2" ($\Gamma_1 С \Gamma_2$), когда место артикуляции при приходе на смычку противоречит месту артикуляции при уходе со смычки, принимая во внимание длительность смычки и мышечные усилия, требуемые для подобной смены места артикуляции.

Решение обратной задачи для взрывных согласных с использованием кодовой книги состоит в определении таких управлений артикуляторами, чтобы на заданных временных интервалах (или в моменты времени) невязка между вычисленными и измеренными акустическими параметрами была бы в пределах погрешности измерений. Это своеобразная задача интерполяции в неравномерных узлах, и только динамические свойства артикуляторов могут помочь в ее решении.

Аналогично формируется кодовая книга для назальных с той разницей, что при размыкании назальной смычки взрыв либо не образуется, либо чрезвычайно слаб. Поэтому вместо отсчета спектральных характеристик взрыва используются формантные частоты в начале переходного процесса гласного. Фрикативные согласные в сочетании с гласными также могут демонстрировать формантные переходы. Поэтому, помимо средних частотных характеристик спектра, в кодовой книге представлены и ячейки с формантными переходами.

5. Управления

Динамика каждого артикуляторного параметра может быть описана обыкновенным дифференциальным уравнением второго порядка

$$m\ddot{y} + r\dot{y} + cy = G(t),$$

где y – артикуляторный параметр, m, r, c – масса, коэффициент вязкого трения и упругое сопротивление артикулятора и связанных с ним тканей, G – сила мышечного сокращения. Это следует из того, что артикулятор либо является системой с сосредоточенными параметрами, либо

может быть сведен к такой системе [14]. В статье [29] было показано, что последовательность команд на мышечные сокращения можно достаточно хорошо аппроксимировать кусочно-линейными функциями, $G=a+bt$.

Динамические параметры артикуляторов приведены в Табл. 2. Диапазон изменения коэффициента a определяется по диапазону изменений артикуляторного параметра, тогда как скорость изменения управления b должна определяться экспериментально. Каноническая форма уравнения второго порядка есть

$$\ddot{y} + 2g\dot{y} + \omega^2 y = f(t),$$

где $g=r/(2m)$, $\omega^2=c/m$, $f=G/m$.

Табл. 2. Диапазон артикуляторных параметров и динамические параметры.

	параметр	min/max (см)	ω (рад/сек)	вес, (г)/ масса ($гсек^2/см$)	жесткость (г/см)	max ***a*** +/- (г)
1	длина губ	3.8/4 (4)	100	10/0.01	100	15/0
2	верт. см. губ	-0.5/0.5	100	10/0.01	100	50/50
3	угол поворота н.ч.	0/0.13 (рад)	40	200/0.2	300	20/20 $l_{пл}$=2см
4	гориз. смещение н.ч.	-0.5/0.8	32	200/0.2	200	160/100
5	угол поворота языка	0/0.3 (рад)	60	55/0.056	200	30/0 $l_{пл}$=2см
6	поперечная деформ языка	0/0.5	42	55/0.056	100	50/0
7	коэф. для конч. яз.	0/0.5	80	16/0.016	100	100/0
8	нижний отдел *genioglos*	-	120	18/0.019	70	30/0
9	средний отдел *genioglos*	-	120	18/0.019	70	30/0
10	верхний отдел *genioglos*	-	120	18/0.019	70	30/0
11	угол небной занавески	0.4/0 (рад)	60	5/0.005	20	40/0 l=2см
12	X корня языка	1.0/2.7 (1.8)	30	55/0.056	50	45/40
13	Y корня языка	4.1/7.1 (5.6)	30	55/0.056	50	75/75
14	средний сжиматель глот	-2/2	60	10/0.01	40	2/0
15	верхний сжиматель глот	-2/2	60	10/0.01	40	2/0
16	высота гортани	-0.5/2.5	30	20/0.02	20	50/10
17	площадь гол. щели	0/0.4 ($см^2$)	60	3/0.003	10	4/4

Как описано в работе [1], параметры управлений находятся по следующей двухэтапной схеме. Первый этап состоит в определении значений артикуляторных параметров как функций времени с использованием акустических данных и опорных точек. Этот этап представляет собой некорректную обратную задачу. Второй этап – определение параметров управлений по найденным временным трекам артикуляторных параметров, оказывается корректной обратной задачей. Остановимся на ее постановке.

Прямая задача вычисления артикуляторного трека $y = y(t)$ по кусочно-линейным управлениям заключается в решении краевой задачи

$$\ddot{y} + 2g\dot{y} + \omega^2 y = A + Bt,\ 0 < t < T,$$
$$y(0) = y^{(0)}, y(T) = y^{(T)},$$

с данными начальными и конечными значениями $y(0) = y^{(0)}$, $y(T) = y^{(T)}$ и с заданным линейным управлением на рассматриваемом отрезке времени $[0,T]$, где коэффициенты A, B этого управления не меняются. Эта краевая задача решается аналитически: $y(t) = y_0(t) + Ay_1(t) + By_2(t),\ 0 \le t \le T,$ где

$$y_0(t) = y^{(0)} e^{-gt} \frac{\mathrm{sh}\,\Omega(T-t)}{\mathrm{sh}\,\Omega T} + y^{(T)} e^{g(T-t)} \frac{\mathrm{sh}\,\Omega t}{\mathrm{sh}\,\Omega T},$$
$$y_1(t) = \frac{1}{\omega^2}\left(1 - e^{-gt} \frac{\mathrm{sh}\,\Omega(T-t)}{\mathrm{sh}\,\Omega T} - e^{g(T-t)} \frac{\mathrm{sh}\,\Omega t}{\mathrm{sh}\,\Omega T}\right),$$
$$y_2(t) = \frac{1}{\omega^4}\left(2ge^{-gt} \frac{\mathrm{sh}\,\Omega(T-t)}{\mathrm{sh}\,\Omega T} - e^{g(T-t)} \frac{\mathrm{sh}\,\Omega t}{\mathrm{sh}\,\Omega T}\left(T\omega^2 - 2g\right) + t\omega^2 - 2g\right).$$

Здесь $\Omega = \sqrt{g^2 - \omega^2}$.

Обратная задача нахождения коэффициентов A, B по треку артикулятора $y_{\mathrm{exp}er}(t_k)$ в точках t_k, $k = 1,...,N$, $t_1 = 0$, $t_N = T$ сводится к переопределенной системе линейных уравнений

$$Ay_1(t_k) + By_2(t_k) = d(t_k),\ k = 1,...,N,$$

где $d(t_k) = y_{\mathrm{exp}er}(t_k) - y_0(t_k)$. В матричной форме система имеет вид

$$\mathbf{C}\begin{bmatrix} A \\ B \end{bmatrix} = \mathbf{d}\ , \qquad (1)$$

с матрицей

$$\mathbf{C} = \left[\mathbf{y_1}\ \mathbf{y_2}\right],\ \mathbf{y_1} = [y_1(t_1)\, y_1(t_2) \ldots y_1(t_N)]^T,\ \mathbf{y_2} = [y_2(t_1)\, y_2(t_2) \ldots y_2(t_N)]^T,$$

и правой частью $\mathbf{d} = [d(t_1)\, d(t_2) \ldots d(t_N)]^T$. Здесь символ *T* обозначает транспонирование матрицы. Решение этой системы в смысле метода наименьших квадратов может быть неединственным. Поэтому мы будем искать так называемое нормальное псевдорешение, то есть вектор, минимизирующий невязку системы (1) и имеющий минимальную евклидову норму. Нормальное псевдорешение выражается в форме

$$\begin{bmatrix} A \\ B \end{bmatrix} = \mathbf{C}^{+}\mathbf{d}\ . \qquad (2)$$

Здесь $\mathbf{C}^{+}$ - псевдообратная матрица для матрицы $\mathbf{C}$. В случае невырожденной квадратной матрицы $\mathbf{C}$ псевдообратная матрица совпадает с обратной. Методы устойчивого вычисления псевдообратной матрицы хорошо известны. Их обзор приведен, например, в [30].

Решение (2) системы (1) устойчиво по отношению к возмущениям данных $\mathbf{d}$. В нашем случае псевдообратная матрица вычислялась с помощью встроенной функции пакета MATLAB.

Оценка параметров кусочно-линейных управлений для каждого *k*-го артикуляторного параметра выполняется на последовательности отрезков времени $T_i^{(k)}$, i=1, … , $N^{(k)}$. Алгоритм автоматического разбиения траектории артикуляторного параметра на эти отрезки описан в [31].

На Рис. 9 показан результат вычисления кусочно-линейных команд для словосочетания “T*he other one is too big*” как решение обратной задачи с входными данными, включающими не только акустические параметры звуков речи, но и траектории опорных точек внутри речевого тракта.

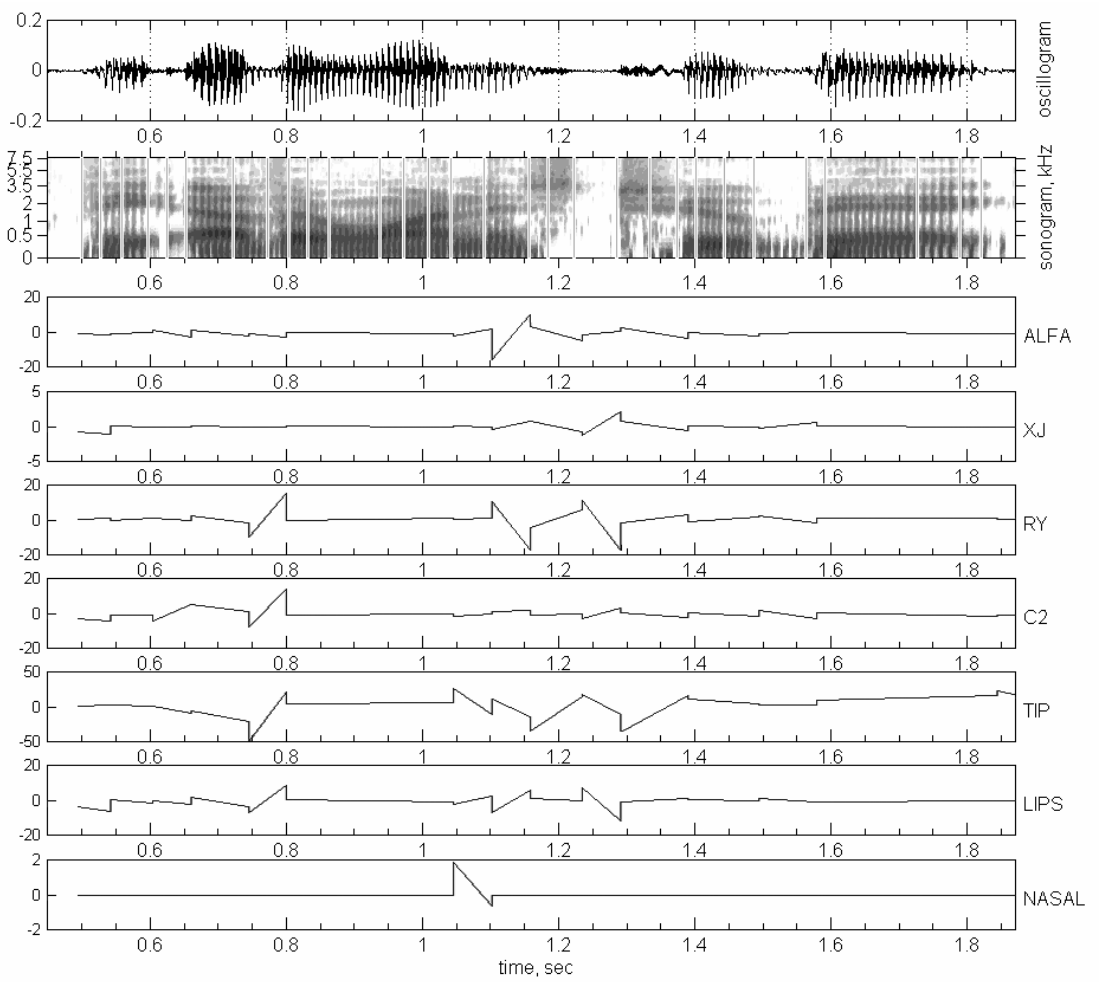

Рис. 9. Осциллограмма звукового давления, сонограмма и управления по некоторым артикуляторным параметрам для фразы "*The other one is too big*". По осям ординат – мышечные усилия в граммах.

Разборчивость и натуральность этой фразы, синтезированной по вычисленным командам управления, можно оценить, прослушивая ее в сравнении с оригинальным произнесением:

The other one is too big - оригинальное

The other one is too big – ресинтез

6. Обсуждение

Эффективность использования кодовой книги проверяется в условиях, когда для решения обратной задачи доступны только акустические данные. В [11] описывались результаты решения обратной задачи для гласных с использованием кодовой книги, сформированной путем вариации параметров модели речеобразования, обеспечивающей достаточно натуральное качество синтеза. Было установлено, что точность восстановления формантных частот гласных звуков составляет 8.4% для F_1, 29.7% для F_2, 12.8% для F_3. Точность аппроксимации формантных частот гласных английского языка с использованием кодовой книги, созданной описанным в настоящей работе способом, оказывается значительно выше. Погрешности не превышают 1.8% для F_1, 1.6% для F_2, 1.1% для F_3, что сопоставимо с погрешностью, полученной при решении обратной задачи с использованием артикуляторных данных: 3.7% для F_1, 3.8% для F_2 и 2.6% для F_3. Аналогичная проверка качества решения обратной задачи для фрикативных показала, что точность аппроксимации характерных частот спектра фрикативных с использованием артикуляторных данных близка к 8.5%, а при использовании только акустических параметров - около 13.4%.

Обучение кодовой книги для многих дикторов напоминает процедуру обучения систем автоматического распознавания речи. Наши эксперименты показывают возможность решения речевой обратной задачи с точностью, достаточной для использования в задачах распознавания, синтеза и сжатия речи. Но существует и физиологический аспект этой проблемы. Если, как это

следует из данных о естественных и искусственных нарушениях процессов речеобразования и восприятия речи, а также гипотезы внутренней модели, обратная задача решается человеком и в процессах управления собственной речью, и при распознавании речи других людей, то как могла бы формироваться кодовая книга в системе управления артикуляцией и восприятия речи?

Человек может построить кодовую книгу для своего речевого тракта, поскольку система управления артикуляцией знает об артикуляторных параметрах, а слуховая система доставляет информацию об акустических параметрах. Такая кодовая книга послужит основой для формирования кодовых книг, т.е. моделей речевых трактов других людей. Во-первых, в процессе роста человека в его памяти могут оставаться кодовые книги для разных размеров тракта. Во-вторых, наблюдения за внешними проявлениями артикуляции могут доставить достаточную информацию для формирования новых кодовых книг. Из ежедневной практики хорошо известно, что визуальная информация облегчает восприятие речи других людей, особенно в неблагоприятных акустических условиях или для иностранного языка [32]. Наблюдение за лицом диктора влияет на восприятие речи, и в случае противоречия между видимой артикуляцией и услышанным звукосочетанием возникают разнообразные эффекты восприятия [33]. Информативность наблюдаемых проявлений артикуляции позволяет общаться глухонемым.

Результаты исследований внешних проявлений аудио-визуальных эффектов подтверждаются исследованием активности мозга. Например, была обнаружена электрическая активность слуховой зоны коры головного мозга слушателя, наблюдающего за артикуляторными движениями диктора, тогда как неречевая мимика диктора такой активности не вызывала [34, 35].

Поэтому, даже не имея доступа к измерениям формы внутренних поверхностей речевого тракта других людей, слушатель мог бы сформировать кодовую книгу, содержащую информацию, которая помогает в оценке моторной компоненты речевого сигнала.

9. Заключение

Кодовая книга для решения речевых обратных задач представляет собой сложную конструкцию, отдельные части которой формируются с учетом особенностей артикуляции и акустических процессов разных типов звуков речи – гласных, назальных, фрикативных и взрывных согласных. Решение обратных задач для разных типов звуков требует разных акустических параметров, причем стационарные и нестационарные сегменты речи представлены в кодовой книге разными структурами. Установлено, что, в отличие от модельных задач со специально подготовленными акустическими данными, формирование кодовой книги для решения обратных задач для произвольной речи произвольного диктора в реальных акустических условиях требует решения задач автоматического вычисления акустических параметров с требуемой точностью и с автоматическим восполнением отсутствующих необходимых данных.

Работа выполнена при поддержке РФФИ, грант №03-01-00116.

Литература

1. A.S.Leonov and V.N.Sorokin. Inverse problem for the vocal tract: Identification of control forces from articulatory movements, Pattern Recognition and Image Analysis, 2000, v. 10, pp. 110-126.
2. А.С.Леонов, И.С.Макаров, В.Н.Сорокин, А.И.Цыплихин, 2003. Артикуляторный ресинтез гласных, Информационные процессы, т. 3, №2, 73-92.
3. А.С.Леонов, И.С.Макаров, В.Н.Сорокин, А.И.Цыплихин. Артикуляторный ресинтез фрикативных, Информационные процессы, 2004, т. 4, №2, с. 141-159.
4. B.S. Atal, J.J. Chang, M.V. Mathews and J.W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique, *J. Acoust. Soc. Am.*, 1978, v. 63, N 5, pp. 1535-1555 .
5. J.N.Larar, J.Schroeter, M.M.Sondhi. Vector-quantization of the articulatory space, IEEE Trans. Acoust., Speech and Signal Processing, 1988, ASSP-36, N 12, pp. 1812-1818 .
6. J.Schroeter, R.Meyer, S.Parthasarathy. Evaluation of improved articulatory codebooks and codebook access distance measures, Proc. Internat. Conf. Acoust. Speech Signal Processing, 1990, pp. 393-396.
7. J. Schroeter and M.M. Sondhi. Dynamic programming search of articulatory codebooks, in: IEEE Proc. Int. Conf. Acoust. Speech Signal Process., 1989, pp. 588-591.
8. J.Schroeter, M.M.Sondhi. Speech coding based on physiological models of speech production, in Advances in Speech Signal Processing, ed. by S.Furui and M.M.Sondhi (The Bartlett Press), 1992, pp. 231-266.
9. J.Schroeter and M.M. Sondhi. Techniques for estimating vocal tract shapes from the speech signal, IEEE Trans on Speech and Audio Proc., 1994, v. 2, N 1, Part 2, pp. 133-150.

10. V.N.Sorokin and A.V. Trushkin. Articulatory-to-acoustic mapping for inverse problem, Speech Communication, 1996, v. 19, N 4, pp. 105-118.
11. V.N.Sorokin, A.S.Leonov and A.V.Trushkin. Estimation of stability and accuracy of inverse problem solution for the vocal tract, Speech Communication, 2000, v. 30, N1, pp. 55-74.
12. V.N.Sorokin. Determination of vocal tract shape for vowels, Speech Communication, 1992, v. 11, N 1, pp. 71-85.
13. V.N.Sorokin. Inverse problem for fricatives, Speech Communication, 1994, v. 14, N 3, pp. 249-262.
14. В.Н.Сорокин. Теория речеобразования, 1985, Радио и связь, М., 313 с.
15. В.Н.Сорокин. Синтез речи. Наука, М., 1992, 392 с.
16. И.С. Макаров, В.Н. Сорокин. Резонансы разветвленного речевого тракта с податливыми стенками, Акустический ж., 2004, т. 50, № 3, с. 389-396.
17. Д.Фланаган. Анализ, синтез и восприятие речи. М., Связь, 1968.
18. D.Kewly-Port, C.S.Watson. Formant-frequency discrimination for isolated English vowels, J. Acoust. Soc. Amer., 1994, v. 95, pp.485-496.
19. В.Н.Сорокин, В.П.Трифоненков. Об автокорреляционном анализе речевого сигнала, Акустический ж., 1996, т. 42, N3, с. 418-425.
20. Vallabha G.K., Tuller B. Systematic errors in the formant analysis of steady-state vowels, Speech Communication, 2002, v. 38, pp. 141–160.
21. В.Н.Сорокин, А.И.Цыплихин. Сегментация и распознавание гласных, Информационные процессы, 2004, т. 4, № 2, с. 202-220.
22. J.Schoentgen, S. Ciocea . Kinematic formant-to-area mapping, Speech Communication, 1997, v. 21, pp. 227-244.
23. S.E.Blumstein, K.N. Stevens. Perceptual invariance and onset spectra for stop consonants in different vowel environments, J. Acoust. Soc. Amer., 1979, v. 67, pp. 648-662.
24. S.Cassidy, J. Harrington. The place of articulation distinction voiced and stops: Evidence from burst spectra and formant transitions, Phonetica, 1995, v. 52, N4, pp. 263-284.
25. D.Kewly-Port. Perception of static and dynamic cues to place of articulation in initial stop consonants, J. Acoust. Soc. Amer., 1983, v. 73, pp. 1779-1992.
26. Z.B.Nosair, S.A.Zahorian. Dynamic spectral shape features as correlates for initial stop consonants, J. Acoust. Soc. Amer., 1991, v. 89, pp. 2978-2991.
27. R.N.Ohde, K.N.Stevens. Effect of burst amplitude on the perception of place of articulation for stops, J. Acoust. Soc. Amer., 1983, v. 74, pp. 706-714.
28. R.Smits, L.Ten Bosch, R.Collier. Evaluation of various sets of acoustical cues for the perception of prevocalic stop consonants, J. Acoust. Soc. Amer., 1996, v. 100, pp. 3852-3864.
29. А.С.Леонов, В.Н.Сорокин. Обратная задача для управления артикуляцией, Доклады Академии Наук, 2000, т. 374, № 6, с. 749-753.
30. А.Н. Тихонов, А.С. Леонов, А.Г. Ягола. Нелинейные некорректные задачи. М., Наука, 1995.
31. A.S.Leonov, V.N.Sorokin. Controls in the internal model: Score reorganization and compensation, Pattern Recognition and Image Analysis, 2004, v.14, N 3, pp. 407-420.
32. W.Y. Sumby, I.Pollack. Visual contribution to speech intelligibility in noise, J. Acoust. Soc. Amer., 1954, v. 26, pp. 212-215.
33. H.McGurk, J.MacDonald . Hearing lips and seeing voices, Nature, 1976, v. 264, pp. 746-748.
34. M.Sams, R.Aulanko, H.Hamalainen, O.Lounasmaa, S.Lu, J.Simola. Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, 1991, v. 127, pp. 141-145.
35. G.Calvert, M.Brammer, E.Bullmore, R.Campbell, S.Williams, P.McGuire, P.Voodruff, S.Iversen, A.David. Activation of auditory cortex during silent lip-reading. *Science*, 1997, v. 276, pp. 593-596.