Информационные процессы, Том 3, №2, стр.73-82

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

АРТИКУЛЯТОРНЫЙ РЕСИНТЕЗ ГЛАСНЫХ

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 19.06.2003

Аннотация - Решалась обратная задача нахождения формы речевого тракта и профиля площадей его поперечного сечения по акустическим и геометрическим данным. Данными служили: первые три резонансные частоты, измеренные в речевом сигнале; соответствующие траектории движений 8 точек на внутренних поверхностях речевого тракта, измеренные на микролучевом рентгеноскопе. Найденное динамическое решение обратной задачи использовалось артикуляторным синтезатором для генерации речевого сигнала. Звучание синтезированных речевых сигналов и их сонограммы оказались весьма близки к оригинальным звукам - 13 гласным и 6 дифтонгам английского языка одного диктора.

1. Введение

Решение обратной задачи для речевого тракта, т.е. восстановления его формы и параметров управления артикуляцией по измеренным акустическим параметрам речевого сигнала, теоретически, создает возможность для построения артикуляторного синтезатора речи, который может использоваться либо для синтеза речи по тексту, либо как оконечный каскад в артикуляторном вокодере. В общем случае, обратная задача для речевого тракта является нелинейной и некорректно поставленной: заданному набору входных данных, как правило, соответствует много формальных решений, большинство из которых неустойчивы по отношению к возмущениям данных. Поэтому для решения указанной обратной задачи необходимо использовать методы и алгоритмы, которые обеспечивают получение физически, физиологически и фонетически приемлемых устойчивых решений. Практическую работоспособность этих алгоритмов можно оценить, применяя процедуру **ресинтеза**: синтезированный по найденному решению речевой сигнал перцептивно должен мало отличается от исходного речевого сигнала, по параметрам которого решалась обратная задача. Субъективно, натуральность речи в значительной степени определяется звучанием гласных звуков. Поэтому в данной работе исследовались методы решения обратной задачи и ресинтеза гласных и дифтонгов.

Теория методов решения нелинейных обратных некорректно поставленных задач развита в [1]. Обзор конкретных методов решения обратных задач для речевого тракта, когда входными данными служат только акустические параметры речевого сигнала, приведен в [2]. Эффективным способом решения обратных задач является вариационный метод (см. [1]). Для речевых обратных задач он сводится к варьированию параметров математической модели речеобразования с целью нахождения глобального минимума заданного критерия оптимальности искомого решения. Критерий оптимальности обычно включает в себя некоторый энергетический критерий и невязку между измеренными и вычисленными параметрами. Минимизация происходит при ограничениях на искомые параметры. Математические модели, связывающие искомые параметры и экспериментальные входные данные, были описаны для всех уровней процесса речеобразования в [3, 4]. Эти модели были использованы ранее при решении обратных задач для стационарных гласных и фрикативных [2, 5, 6], а также при решении некоторых динамических задач [7, 8].

Для эффективного решения задач на условный глобальный минимум в вариационном методе важно иметь “хорошее” начальное приближение. Только в этом случае процесс минимизации может дать необходимое приближенное решение. Для поиска начального приближения в [9] был предложен метод "кодовой книги". Первоначально кодовая книга создавалась с использованием артикуляторно-акустических моделей речеобразования путем перебора всевозможных сочетаний артикуляторных параметров и вычисления соответствующих резонансных частот речевого тракта. Тогда каждому вектору акустических параметров речевого сигнала можно поставить в соответствие некоторое множество векторов артикуляторных параметров, которые при решении прямой задачи дают акустические параметры, близкие к заданным. Артикуляторные параметры из такого множества и

служат начальными приближениями при решении обратной задачи для реального речевого сигнала. В действительности, возможны не все сочетания артикуляторных параметров. Число ячеек в кодовой книге можно уменьшить почти на два порядка, если отсчеты акустических и артикуляторных параметров берутся на их траекториях при синтезе слитной речи [10]. Другой подход к построению кодовой книги состоит в сопоставлении артикуляторных параметров реальных измеренных форм речевого тракта и соответствующих измеренных параметров речевого сигнала. Реализация этого подхода была невозможна до тех пор, пока не появились достаточно представительные базы экспериментальных данных о форме речевого тракта для различных звуков. В настоящей работе был принят именно такой способ формирования кодовой книги. Детали описаны в [11].

2. Экспериментальные данные

Экспериментальной основой исследования послужила база данных, сформированная по измерениям на микролучевом рентгеноскопе [12]. В ней представлены анатомические параметры 47 дикторов, запись речевого сигнала и синхронные измерения координат 8 точек (точек наблюдения) внутри речевого тракта при произнесении разнообразных текстов. На рис. 1 приведена характерная форма речевого тракта, реконструированная по анатомическим параметрам одного из дикторов, и характерные положения точек наблюдения. В наших экспериментах использовались данные из базы для 13 изолированных гласных, произносимых как в словах: *er* [dirt], *uh* [but], *uu* [boot], *ay* [date], *aw* [bought], *oh* [boat], *ih* [bit], *ah* [hot], *ee* [beet], *eh* [bet], *oo* [foot], *ae* [bat], и 6 дифтонгов: *eeoo* [iu], *eeah* [ia], *ooah* [ua], *ahoo* [au], *ahee* [ai], *ooee* [ui]. Обозначения гласных соответствуют классификации, принятой в разметке речевых сигналов на фонетические сегменты американского английского языка.

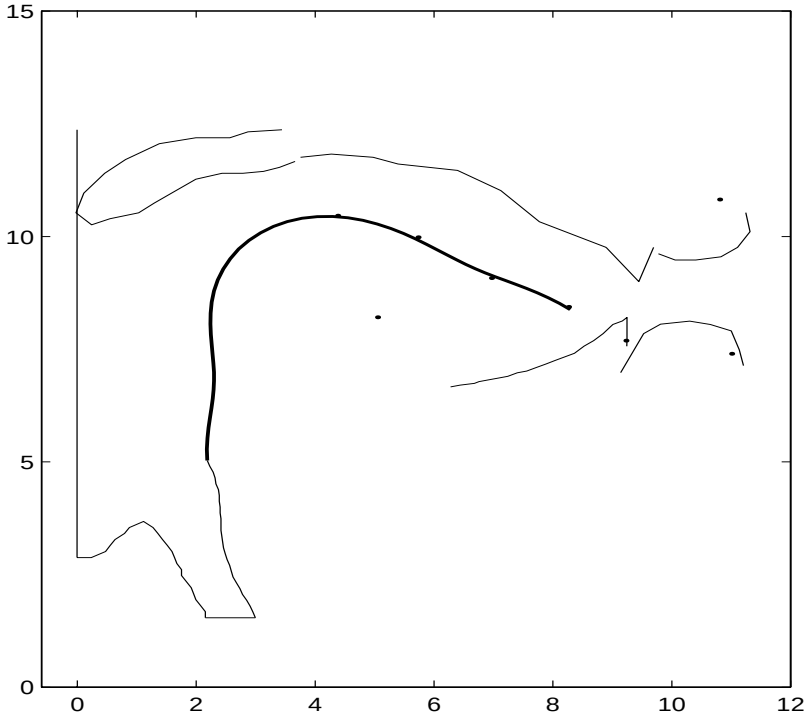

Рис. 1. Расположение точек измерения внутри речевого тракта.

По акустическим данным из базы находились спектры соответствующих гласных и дифтонгов. Спектры вычислялись в скользящем окне Хемминга на интервале в 1024 отсчета (частота отсчетов речевого сигнала – 21.739 кГц). Частоты максимумов спектра, примерно соответствующие резонансным частотам речевого тракта, измерялись для каждого импульса возбуждения голосового источника. Временные треки этих оценок формантных частот, а также треки 8 точек речевого тракта с отсчетами через 7 мс использовались как данные для решения обратной задачи.

3. Математические модели артикуляции и акустики

Для решения обратной задачи и синтеза речи использовалась 15-параметрическая артикуляторная модель [3, 4]. Она использует следующие артикуляторные параметры: высоту голосовой щели, координаты корня языка, координаты кончика языка, угол поворота нижней челюсти, горизонтальное смещение точки вращения нижней челюсти, 5 коэффициентов при собственных функциях упругих деформаций языка, высоту нижней губы, и 2 коэффициента при

собственных функциях, описывающих изменение ширины глотки. Кроме того, из базы данных для каждого диктора выбирались его анатомические параметры. К их числу относятся форма свода твердого неба, размеры верхней и нижней челюсти, параметры ширины глотки в нейтральном состоянии, параметры импеданса стенок речевого тракта, а также длина и площадь поперечного сечения грушевидных полостей, разветвляющих речевой тракт на уровне входа в пищевод.

Артикуляторная модель вместе с анатомическими параметрами, измеренными для данного диктора, позволяет вычислить форму речевого тракта и найти распределение площадей $S(x,t)$ его поперечных сечений вдоль его средней линии [13]. По известной функции $S(x,t)$ можно найти резонансные частоты речевого тракта, считая эту функцию медленно меняющейся между временными отсчетами экспериментальных данных: $S(x,t)\approx S(x)$. Резонансные частоты находятся из решения спектральной задачи для волнового уравнения, описывающего акустические процессы в речевом тракте

$$\frac{\partial}{\partial x}(S(x)\frac{\partial p}{\partial x}) = \frac{S(x)}{c_0^2(x)}\frac{\partial^2 p}{\partial t^2},$$

с соответствующими граничными условиями. Здесь p - звуковое давление, $S(x)$ - площадь поперечного сечения тракта, x - координата вдоль средней линии тракта, t - время, c_0 - скорость звука в воздухе. Для учета податливости стенок речевого тракта и его разветвления вычисление собственных частот удобно выполнять с использованием метода длинных линий [14]. Согласно этому методу, речевой тракт аппроксимируется последовательностью N цилиндрических секций длиной Δl. Входной акустический импеданс Z_g речевого тракта со стороны гортани в предположении абсолютной жесткости стенок тракта и отсутствия разветвлений вычисляется из

$$Z_g(z) = Z_{0N}\frac{Q_N + P_N}{Q_N - P_N}, \quad (1)$$

где

$$Q_i = Q_{i-1} + r_{i-1}z^{-1}P_{i-1},$$
$$P_i = r_{i-1}Q_{i-1} + z^{-1}P_{i-1},$$
$$Q_1 = Z_L + Z_{01},$$
$$P_1 = Z_L - Z_{01}$$

Коэффициенты отражения определяются как $r_i = \frac{S_{i+1} - S_i}{S_{i+1} + S_i}$, а индекс i пробегает все значения от 2 до N. Здесь $Z_{0i} = \frac{\rho c_0}{S_i}$ - характеристический импеданс i-й секции, ρ – плотность воздуха (0.00114 г/см^3), S_i - площадь поперечного сечения i-й секции, $z = \exp(j2\omega\Delta l / c_0)$, j –мнимая единица, ω круговая частота (рад/с), Z_L - импеданс излучения через губы, определяемый соотношением [15]

$$Z_L = \frac{1 - z^{-1}}{\frac{2}{R} + 0.7(1 - z^{-1})}, \quad (2)$$

R – эффективный радиус губного отверстия. При разветвлении в речевом тракте в m-й секции входной акустический импеданс со стороны гортани также вычисляется по формуле (1) с той разницей, что Z_L определяется соотношением (3), а не (2):

$$Z_L = \frac{Z_m Z_r}{Z_m + Z_r}, \quad (3)$$

а индекс i пробегает значения от m до N (а не от 1 до N, как в соотношениях (1)). Здесь Z_m- входной акустический импеданс в m-ю секцию, Z_r - входной акустический импеданс в разветвляющую трубу

со стороны речевого тракта. Z_m вычисляется с помощью соотношений (1), (2), где индекс i пробегает значения от *1* до *m*. Для вычисления Z_r необходимо предварительно аппроксимировать разветвляющую трубу последовательностью M цилиндрических труб длины Δl. В этом случае Z_r также вычисляется с помощью формул (1), где S_i - площадь поперечного сечения i-й секции разветвляющей трубы, $i = 1, \ldots, M$, Z_L - импеданс нагрузки разветвляющей трубы (для грушевидных разветвлений $Z_L = \infty$).

Если входной акустический импеданс Z_g в речевой тракт с абсолютно жесткими стенками известен, то входной акустический импеданс $\overline{Z}_g$ в речевом тракте с податливыми стенками определяется, по [16], как

$$\overline{Z}_g = Z_g(\sigma)\sqrt{\frac{j\omega}{j\omega+\beta}},$$

$$\sigma = \sqrt{j\omega\beta - \omega^2},$$

$$\beta = \frac{90000\pi^2 j\omega}{j\omega(1+\frac{R_w^{'}}{L_w^{'}}) + \frac{1}{L_w^{'}C_w^{'}}}$$

Здесь $R_w^{'}, L_w^{'}, C_w^{'}$ - соответственно, активные потери на стенках тракта, масса и упругость стенок тракта. Резонансные частоты речевого тракта $\{F_i\}$ определяются как мнимые части полюсов входного акустического импеданса $\overline{Z}_g$. Ширина резонансов $\{B_i\}$ описывается эмпирически найденными соотношениями [17]

$$B_i = 15\left(\frac{500}{F_i}\right)^2 + 20\sqrt{\left(\frac{F_i}{500}\right)} + 2.8\left(\frac{F_i}{500}\right)^2$$

4. Постановка обратной задачи и алгоритм ее решения

Обозначим вектор артикуляторных параметров, определяющих форму речевого тракта, как $u=(u_1,u_2,\ldots, u_n)$. Этот вектор в силу артикуляторной и акустической модели, описанной выше, позволяет вычислить вектор наблюдаемых величин $v=(v_1,v_2,\ldots, v_m)$, то есть совокупность координат отслеживаемых восьми точек речевого тракта вместе с первыми тремя резонансными частотами. Связь вектора артикуляторных параметров и вектора наблюдаемых величин может быть записана в виде $v=A(u)$, где нелинейный непрерывный оператор A математически представляет артикуляторную и акустическую модель и реализуется с помощью специальной вычислительной процедуры. Для заданного динамического трека $u=u(t)$ артикуляторных параметров можно таким образом вычислить соответствующий трек наблюдаемых величин $v=v(t)$. Решение обратной задачи: “от наблюдаемых величин к артикуляторным параметрам” соответствует обращению оператора A. Именно эта процедура обращения может давать не единственное и неустойчивое решение. Применяя вариационный метод решения обратной задачи, будем по динамическим трекам наблюдаемых величин $v(t)$ на заданном временном отрезке $t\in[t_{min}, t_{max}]$ искать такой вектор артикуляторных параметров $\overline{u}(t)$, для которого

$$\Omega[\overline{u}(t)] = \min\{\Omega[u(t)]: \|A(u) - v\| = 0\}. \quad (4)$$

Это означает, что вектор $\overline{u}(t)$ - это то из решений уравнения $v=A(u)$, для которого минимален выбранный критерий оптимальности $\Omega[u]$. Здесь норма $\|\cdot\|$ определяется спецификой задачи, а также выбранным критерием $\Omega[u]$.

На практике вектор $v(t)$ – данные для решения обратной задачи – известен приближенно вследствие ошибок его измерения или вычисления. Поэтому мы будем считать, что вместо $v(t)$ в нашем распоряжении имеется его реализация $v_\delta(t)$, найденная с некоторой известной

среднеквадратичной точностью δ такой, что выполнено неравенство $\|v - v_\delta\| \le \delta \|v_\delta\|$. В этом случае вместо решения экстремальной задачи (4) обычно решается ее приближенный аналог: найти такой вектор артикуляторных параметров $u_\delta(t)$, что он удовлетворяет некоторым априорным ограничениям, задаваемым множеством U, и для которого

$$\Omega[u_\delta(t)] = \min\{\Omega[u(t)]: u = u(t) \in U, \ \|A(u) - v_\delta\| \le \delta \|v_\delta\|\}. \quad (5)$$

Задача (5) означает следующее. Рассматриваются все допустимые артикуляторные параметры $u(t)$: $u(t) \in U$. Среди них выбираются те, которые обеспечивают "близость" (с точностью δ) вычисляемых с помощью модели наблюдаемых величин $v = A(u)$ и приближенных данных $v_\delta(t)$: $\|A(u) - v_\delta\| \le \delta \|v_\delta\|$. Затем из всех таких артикуляторных параметров выбирается такой набор $u_\delta(t)$, для которого минимален критерий оптимальности $\Omega[u]$. Из теории методов решения некорректных задач [1] известно, что при выполнении определенных математических свойств функционала $\Omega[u]$ и оператора A экстремальная задача (5) задача имеет единственное решение $u_\delta(t)$, и оно "близко" к единственному точному решению $\overline{u}(t)$ задачи (4) при "малых" уровнях погрешности данных δ. Таким образом, артикуляторные параметры $u_\delta(t)$, найденные как решение задачи (5), оказываются единственным и устойчивым приближенным решением рассматриваемой обратной задачи.

Существенную роль в такой вариационной процедуре решения обратной задачи играет критерий оптимальности $\Omega[u]$. Он должен с одной стороны удовлетворять необходимым математическим требованиям, а с другой стороны иметь ясную физическую или физиологическую интерпретацию, совместимую с требованиями артикуляторной и акустической модели. Не меньшую роль играет и выбор нормы в задачах (4), (5). В данной работе использовался критерий оптимальности вида

$$\Omega_{PW}[u(t)] = \sum_{k=1}^{n} c_k [x_k(t) - x_k^{(0)}]^2. \quad (6)$$

Здесь $x_k^{(0)}$ - координаты вектора артикуляторных параметров в нейтральном состоянии, а c_k – константы жесткости тканей, связанных с *k*-м артикулятором. Этот критерий имеет физическую интерпретацию: функционал (6) пропорционален средней работе упругих сил в рассматриваемой системе артикуляторов при ее эволюции в течение некоторого малого промежутка времени. Обычно этот промежуток определяется шагом временной дискретизации измерений. Обсуждение и физиологическое обоснование этого критерия, а также некоторых других, дано в [7].

Вид нормы в задачах (4) и (5) определяется структурой наблюдаемых (измеряемых) величин. В данном случае, они разнородны: это и координаты точек наблюдения, и резонансные частоты тракта. Поэтому для адекватного учета вклада каждой из этих величин в невязку $\|A(u) - v_\delta\|$ операторного уравнения использовалась специальная взвешенная норма. Выбор такой нормы и весов определялся предыдущим опытом решения обратных задач "от треков наблюдаемых точек тракта к артикуляторным параметрам" и "от формантных частот к артикуляторным параметрам".

Ограничения U в задаче (5) определяются геометрическими и физиологическими особенностями речевого тракта. Эти ограничения в основном имеют вид неравенств и описывают такие стандартные требования как "язык находится не выше верхнего нёба", "кривизна поверхности языка физиологически допустима", "поверхность зубов не пересекает поверхности языка" и т.п. Правильный выбор ограничения в задаче (5) делает в ряде случаев эту задачу оптимизации достаточно хорошо обусловленной. Задача на условный экстремум (5) решалась численно по схеме из [1]. При этом использовалась процедура последовательной квадратичной аппроксимации этой задачи с применением квазиньютоновского алгоритма минимизации.

Отметим, что математические свойства критерия оптимальности (6) гарантируют следующее важное обстоятельство. Найденные при решении задачи (5) артикуляторные параметры $u_\delta(t)$ адекватно симулируют в рассматриваемой модели наблюдаемые величины: $\|A(u_\delta) - v_\delta\| \to 0$ при $\overline{u}(t)$ [1]. Таким образом, вычисляемые по оптимальным артикуляторным параметрам координаты точек наблюдения в тракте и его главные формантные частоты будут близки к измеряемым, если погрешности данных "малы". Это можно проиллюстрировать некоторыми результатами расчетов (см. Таблицу).

Таблица. Среднеквадратические ошибки решения обратной задачи для координат измеренных точек внутри речевого тракта и первых трех резонансных частот.

Гласные в словах и дифтонги	Средние ошибки по точкам (%)	Средние ошибки по частотам (%)		
		F1	F2	F3
bat	2.7	7.8	7.3	2.5
beet	2.5	2.9	2.6	5.0
bet	2.1	6.6	7.4	2.5
bit	2.6	3.9	4.4	2.6
boat	2.5	2.7	3.6	2.6
boot	3.2	2.8	2.9	4.2
bought	3.7	4.4	2.1	3.0
but	2.7	1.2	1.4	1.1
date	2.6	7.8	9.0	3.3
dirty	4.1	1.7	1.5	3.4
foot	2.9	1.8	1.2	1.2
hot	3.6	4.2	1.4	0.8
ai	2.7	5.0	3.6	3.5
ia	3.4	3.9	4.2	4.0
iu	2.6	5.2	3.9	4.7
ui	3.0	3.5	4.3	3.5
ua	3.2	2.0	3.1	2.3
Среднее	2.8	3.7	3.8	2.6

5. Синтез речи

В процессе синтеза речевого сигнала сначала по вычисленной передаточной функции речевого тракта определялась его импульсная характеристика, а затем вычислялась свертка этой импульсной характеристики с импульсом голосового источника. Передаточная функция речевого тракта вычислялась на каждом периоде основного тона либо по каскадной схеме, либо по параллельной схеме. Согласно [17], амплитуды резонансов для гласных звуков (хотя и не для всех) в каскадной схеме устанавливаются автоматически похожими на амплитудные соотношения в реальной речи, и раздельное вычисление амплитуд для разных резонансов не требуется. В параллельной схеме форма амплитудно-частотной характеристики речевого тракта управляется как резонансными частотами, так и их амплитудами.

В каскадной схеме передаточная функция $T_i(j\omega)$ определяется как

$$T_i(j\omega) = \prod_{k=1}^{5} \frac{1 - 2\exp(-\pi \frac{B_k}{F_d})\cos(\frac{\omega_k}{F_d}) + \exp(-2\pi \frac{B_k}{F_d})}{1 - 2\exp(-\pi \frac{B_k}{F_d})\cos(\frac{\omega_k}{F_d})\exp(-j\frac{2\omega\Delta l}{c_0}) + \exp(-2\pi \frac{B_k}{F_d})\exp(-j\frac{4\omega\Delta l}{c_0})},$$

где F_d - частота дискретизации синтезированного речевого сигнала, $\omega_k = 2\pi F_k$, где $\{F_k\}$ и $\{B_k\}$ - частота и ширина полосы k-го резонанса (в Гц), вычисленные в момент возбуждения речевого тракта i-ым импульсом голосового источника.

В параллельной схеме $T_i(j\omega)$ определяется как

$$T_i(j\omega) = \sum_{k=1}^{5} \frac{A_k}{1 - 2\exp(-\pi \frac{B_k}{F_d})\cos(\frac{\omega_k}{F_d})\exp(-j\frac{2\omega\Delta l}{c_0}) + \exp(-2\pi \frac{B_k}{F_d})\exp(-j\frac{4\omega\Delta l}{c_0})},$$

A_k - амплитуды резонансов, подлежащие управлению.

Качество синтезированного сигнала в значительной степени зависит от источника голосового возбуждения. Для синтеза гласных использовалась модификация модели голосового источника,

описанного в [4]. В этой модели выполняется точное решение уравнения аэродинамического потока через голосовую щель

$$\rho_0 h_\Gamma \left(1+\frac{S_0}{2S_\Gamma}\right)w' + \left(k_{mp}h_\Gamma - \rho_0 h_\Gamma \frac{S_0 S'_\Gamma}{2S_\Gamma^2}\right)w + \frac{\rho_0 c_x}{2S_\Gamma}w^2 = \Delta p S_\Gamma .$$

Здесь w –объемная скорость потока, h_Γ – глубина голосовой щели, S_Γ – площадь голосовой щели, S_0 – площадь трахеи на входе в голосовую щель, c_x – коэффициент динамического сопротивления, Δp – перепад давления на голосовой щели, k_{mp} – коэффициент вязкого трения. Решение этого уравнения в конечных разностях есть

$$w(t+\Delta t) = \frac{\sqrt{1+4a_1a_2}-1}{2a_2},$$

где

$$a_1 = w(t) + \alpha\left[\beta\Delta p(t+\Delta t)S_\Gamma(t+\Delta t) - w(t)\right],\ a_2 = \frac{\alpha\beta c_x\rho_0}{2S_\Gamma(t+\Delta t)},$$

$$\alpha = 1 - e^{-\frac{\Delta t}{\beta\cdot\rho_0 h_\Gamma\left(1+\frac{S_0}{2S_\Gamma}\right)}},\ \beta = k_{mp} - \frac{\rho_0 h S_0 S'_\Gamma}{2S_\Gamma^2},\ F = \Delta p S_\Gamma - \frac{c_x\rho_0}{2}\frac{w^2}{S_\Gamma}.$$

Модель имеет возможность управления длительностью периода возбуждения источника, а также относительными длительностями интервалов открытой и закрытой голосовой щели. Закон изменения площади поперечного сечения голосовой щели описывается некоторой феноменологической формулой

$$S_\Gamma(t) = S_{open} + \begin{cases} 0{,}5S^*\left[1-\cos\dfrac{\pi t}{t_1}\right], 0 \le t < t_1, \\ S^*\cos\left[\dfrac{\pi(t-t_1)}{2(t_2-t_1)}\right], t_1 \le t < t_2, \\ 0, t_2 \le t < T. \end{cases}$$

где

$$S^* = \begin{cases} S_{\Gamma\max}\left(\dfrac{S_{open}-2S_{\Gamma\max}}{2S_{\Gamma\max}}\right)^3,\ 0 \le S_{open} < 2S_{\Gamma\max}, \\ 0,\ S_{open} \ge 2S_{\Gamma\max}. \end{cases}$$

$S_{\Gamma\max}$ – амплитуда колебаний площади голосовой щели, T – период импульса, t_1 – длительность интервала открытия голосовой щели, $(t_2 - t_1)$ – длительность интервала закрытия.

Известно, что после начала глухой смычки или в конце гласного перед паузой голосовые складки расходятся и их колебания затухают. Параметр S_{open} характеризует степень раскрытия голосовой щели. При $S_{open} = 0$ коэффициент $S^* = S_{\Gamma\max}$. При увеличении S_{open} коэффициент S^* нелинейно уменьшается и достигает нуля при $S_{open} \ge 2S_{\Gamma\max}$.

Важной характеристикой голосового источника является шум турбулизации потока на выходе из голосовой щели. Этот шум возникает, когда число Рейнольдса Re превосходит критическое значение Re_{cr}. В данной работе этот шум моделировался двумя случайными процессами с нулевым средним и амплитудами $A_1\left(\mathrm{Re}^2 - \mathrm{Re}_{cr}{}^2\right)$ и $A_2\left(\mathrm{Re}^2 - \mathrm{Re}_{cr}{}^2\right)$, причем полагалось, что коэффициент

$A_1 = 10^{-6}$, а $A_2 = 0{,}1A_1$. Для этих двух процессов порождаемые генератором нормально распределенных случайных чисел значения пропускались через фильтр низких частот с частотами среза $F_1 = 0.085w(t)l / S_\Gamma^2(t)$ и $F_2 = 2F_1$ соответственно, l – длина голосовых складок.

В экспериментах по ресинтезу гласных и дифтонгов (по результатам решения обратной задачи) длительность каждого периода основного тона заимствовалась из реального сигнала. Это вполне оправдано, поскольку информация об основном тоне присутствует в речевом сигнале, и для ее получения не нужно решать обратную задачу. С целью субъективной оценки степени сходства исходных звуков речи и результатов синтеза, оба сигнала, разделенные паузой, предъявлялись для прослушивания.

При вычислении передаточной функции по каскадной схеме для некоторых гласных и дифтонгов различие между синтезированным и исходным сигналом было едва различимо. В других случаях разница была заметна, хотя не только фонетические, но и тембральные характеристики были очень похожи. Более объективная оценка качества синтеза может быть получена путем сравнения динамических спектров (сонограмм) синтезированных и оригинальных сигналов (рис. 2, 3, 4). Видно, что частотные характеристики сигналов практически совпадают, тогда как в амплитудах резонансов наблюдаются некоторые различия.

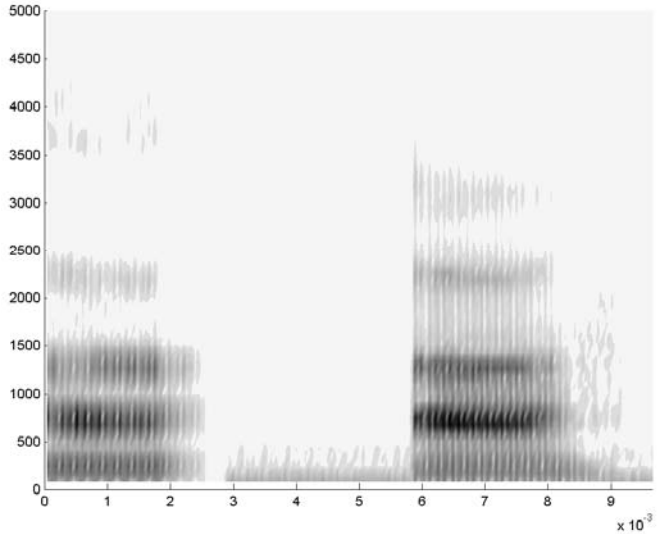

Рис. 2. Сонограммы синтезированного (слева) и оригинального гласного (справа) *er*.

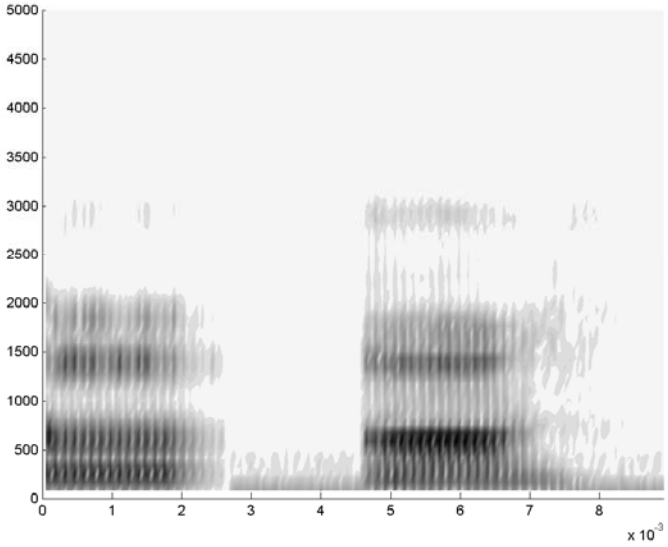

Рис. 3. Сонограммы синтезированного (слева) и оригинального гласного (справа) *uh*.

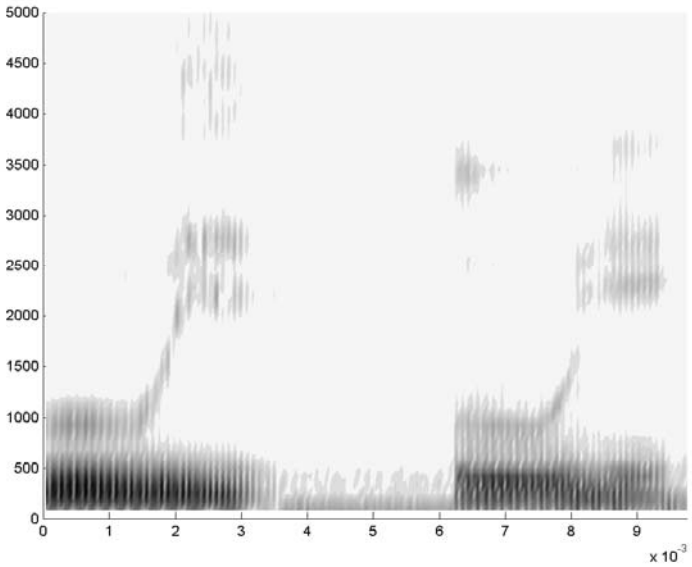

Рис. 4. Сонограммы синтезированного (слева) и оригинального дифтонга (справа) *ui.*

Гласные *ee* и *ae*, которые в каскадной схеме были синтезированы с заметным различием от исходных звуков речи, были синтезированы в параллельной схеме. При этом амплитуды резонансов синтезированных звуков согласовывались с измеренными амплитудами формант (пиков спектра) исходных звуков. Как и следовало ожидать, качество синтеза при этом улучшилось, а вычисленные и измеренные амплитудно-частотные спектры оказались объективно близки (рис. 5, 6). Как видно из этих рисунков, относительные амплитуды резонансов на частотах 755 Гц, 1572 Гц, 2514 Гц, 3962 Гц и 4617 Гц для синтезированного и исходного звука *ae* довольно близки. Разница спектров проявляется в уровне низких частот, что связано с особенностями модели голосового источника.

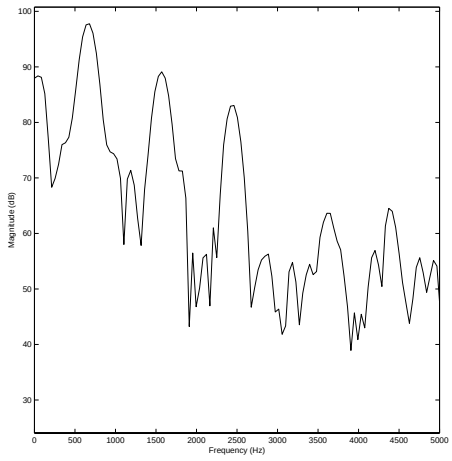

Рис. 5. Спектр синтезированного гласного *ae.*

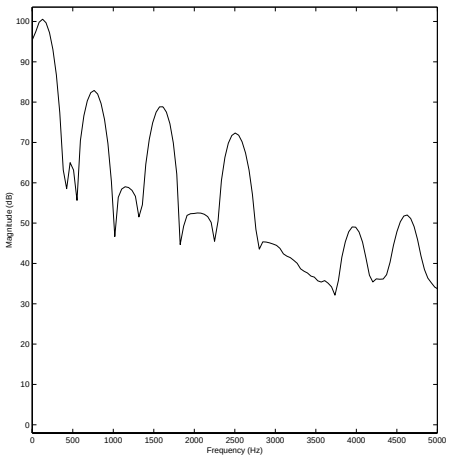

Рис. 6. Спектр оригинального гласного *ae.*

6. Заключение

Гласные и дифтонги, синтезированные по результатам решения обратной задачи относительно формы речевого тракта, весьма близки к оригиналам по субъективной оценке звучания и некоторым спектральным характеристикам. Учитывая, что для голосового источника обратная задача не решалась, и для возбуждения акустических колебаний использовалась довольно грубая модель, результаты ресинтеза следует признать весьма успешными. Таким образом, показана принципиальная возможность использования решений обратной задачи в артикуляторном синтезе речи.

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

Рекомендовано членом редколлегии акад. Н.А.Кузнецовым

Список литературы

1. Тихонов А.Н., Леонов А.С., Ягола А.Г. Нелинейные некорректные задачи. М.: Наука, 1995.
2. Sorokin V.N., Leonov A.S., Trushkin A.V. Estimation of stability and accuracy of inverse problem solution for the vocal tract. Speech Communication, 2000, vol. 30, N1, pp. 55-74.
3. Сорокин В.Н.. Теория речеобразования. М.:Радио и связь, 1985.
4. Сорокин В.Н.. Синтез речи. М.; Наука, 1992.
5. Сорокин В.Н.. Обратная задача для формы речевого тракта. Доклады Академии Наук, 1991, т. 317, N 4, стр. 856-858.
6. Sorokin V.N. Inverse problem for fricatives. Speech Communication, 1994, vol. 14, N 3, pp. 249 - 262.
7. Леонов А.С., Сорокин В.Н. Обратная задача для управления артикуляцией, Доклады Академии Наук, 2000,т. 374, № 6, стр. 749-753.
8. Leonov A.S., Sorokin V.N. Inverse problem for the vocal tract: Identification of control forces from articulatory movements. Pattern Recognition and Image Analysis, 2000, vol. 10, pp. 110-126.
9. Atal B.S., Chang J.J., Mathews M.V. and JTukey.W. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. Journal of Acoustical Society of the America, 1978, vol. 63, N 5, pp. 1535-1555.
10. Sorokin V.N., Trushkin A.V. Articulatory-to-acoustic mapping for inverse problem. Speech Communication, 1996, vol. 19, N 4, pp. 105-118.
11. Леонов А.С., Макаров И.С., Сорокин В.Н. Обучающая фонетическая система, Тезисы 4-й международной конференции “Фонетика сегодня: актуальные проблемы и университетское образование”, М.: Наука, 2003, стр. 79-80.
12. Westbury J. X-ray Microbeam Speech Production Database. User’s Handbook, Version 1, 1994.
13. Макаров И.С., Badin P., Сорокин В.Н. Трехмерная модель речевого тракта и алгоритм вычисления площадей поперечных сечений, Труды Международного семинара Диалог 2002, М.:Наука, 2002, стр. 352-359.
14. Фланаган Дж. Анализ, синтез и восприятие речи. М: Связь, 1968.
15. Rubin P., Baer T., Mermelstein P. An articulatory synthesizer for perceptual research. Journal of Acoustical Society of the America. 1981, vol.70, N 2, pp. 317-328.
16. Sondhi M.M., Shroeter J. Speech Coding based on Physiological Models of Speech Production, “Advances in Speech Signal Processing”, Eds. S. Furui, M.M. Sondhi, 1991, N 9, pp. 231-267.
17. Ladefoged P. A phonation type synthesizer for use in the field, UCLA Working papers in Phonetics 88, September, 1994, pp. 1-11.
18. Klatt D. Software for a cascade/parallel synthesizer . Journal of Acoustical Society of the America. 1980, vol. 67, N3 , pp. 971-995.