Фундаментальные исследования речи и прикладные задачи речевых технологий

В.Н. Сорокин,
доктор физико-математических наук

При вычислении акустических характеристик речевого тракта необходимо учитывать податливость стенок, разветвление в области грушевидных полостей, а также управляемость шириной глотки. Теория внутренней модели подкрепляется экспериментально доказанной возможностью решения обратных задач относительно формы речевого тракта, положения артикуляторов и команд управления с точностью, сравнимой с точностью измерения. На основе математических моделей речевого сигнала показана принципиальная возможность его сжатия до скоростей менее 2 Кб/с с сохранением всех объективных и субъективных показателей качества голоса диктора. Описывается система верификации диктора, обеспечивающая, в среднем, суммарную ошибку пропуска самозванца и отказа законному пользователю ниже 0.01% для подавляющего большинства дикторов.

1. Введение

Речевые исследования носят двойственный характер. С одной стороны, это традиционные фундаментальные исследования в области наук о человеке, а, с другой — это разработка решений для прикладных задач — автоматического распознавания речи, синтеза речи по произвольному тексту, идентификации и верификации диктора, сжатия речевого сигнала в каналах связи.

К фундаментальным проблемам относятся: нейрофизиология управления артикуляцией, обучение языку, компенсация и адаптация к помехам артикуляции, связь между артикуляцией и акустикой, механизмы восприятия, распознавания и понимания речи человеком. Прикладные задачи речевой технологии, за исключением задачи сжатия речи, которая, кстати, удовлетворительно решается формальными методами только для достаточно больших скоростей передачи, решаются с довольно скромным успехом, несмотря на полувековую историю прикладных исследований.

Как известно, фундаментальные исследования в какой-то мере самоподдерживающиеся, поскольку новые направления возникают по ходу решения изначально поставленных научных задач. Но, в дополнение к этому механизму, речевые исследования стимулируются и прикладными задачами, для решения которых оказывается необходимым проведение чисто научных исследований. Надо напомнить, что интенсивные исследования речи начались благодаря постановке задачи речевого общения человека с автоматом, т.е. автоматического распознавания и синтеза речи по заданному тексту.

2. Акустика речеобразования

Из теории речеобразования известно, что в диапазоне частот примерно до 4.5 кГц акустическое давление в речевом тракте описывается одномерным волновым уравнением (модифицированным уравнением Вебстера):

$$\frac{\partial^2 P}{\partial t^2} = c_0^2 \frac{1}{S(x)} \frac{\partial}{\partial x}\left(S(x) \frac{\partial P}{\partial x} \right) - 2\nu \frac{\partial P}{\partial t} + F(x,t), \quad 0 < x < l, \; t > 0 \qquad (1)$$

$$\left.\frac{\partial P}{\partial x}\right|_{x=0} = -\dot{q}(t); \quad \left(\frac{\partial P}{\partial x} - bP \right)_{x=l} = 0$$

$$P(x,0) = P_0(x); \quad \frac{\partial P}{\partial t}(x,0) = P_1(x)$$

где $\nu = \frac{r_1}{2\rho}$ — краевой источник возбуждения (голосовой источник), $q(t)$ — начальные профили давления и скорости его изменения в тракте. Здесь x — пространственная координата вдоль средней линии тракта в среднесагиттальной плоскости, t — момент времени, $P(x,t)$ — искомое давление в тракте, $S(x)$ — профиль площадей поперечного сечения вдоль тракта, $F(x,t)$ — плотность распределения источников возбуждения внутри тракта, c_0 — скорость звука в тракте. Такая модель в принципе пригодна не только для гласных звуков речи, но и для фрикативных звуков, источником возбуждения которых служит шум турбулентного потока воздуха.

Решая спектральную задачу для этого уравнения, получаем собственные числа, а с ними и резонансные частоты речевого тракта, которые ассоциируются с формантными частотами, и формальные модели это подтверждают. Однако, до последнего времени не было возможности измерить реальную функцию $S(x)$, поскольку немногочисленные измерения выполнялись на двумерных рентгенограммах. С появлением метода трехмерной томографии (MRI) стало возможным уточнение отношений между площадью поперечного сечения тракта и его акустическими характеристиками.

В работе [1] было показано, что в процессе артикуляции ширина заднего отдела тракта (от входа в пищевод до мягкого неба) активно изменяется, и достаточно точно описывается линейной комбинацией двух собственных векторов. В работе [2] было найдено, что резонансные частоты, вычисленные по экспериментально измеренным площадям поперечных сечений в предположении жестких стенок, во многих случаях весьма сильно отличаются от измеренных резонансных частот. Формантные частоты, вычисленные в предположении абсолютно жестких стенок, значительно разнятся от измеренных формантных частот. Учет податливости стенок приводит к снижению ошибки по первому резонансу F1. Кроме того, было установлено заметное влияние грушевидных пазух в области гортани на F2 и F3.

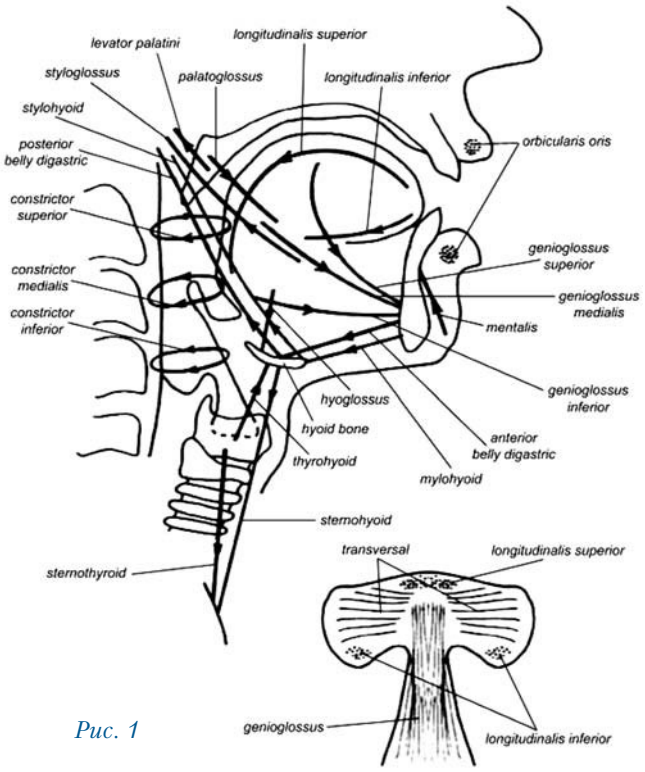

Рис. 1

3. Модель артикуляции

В основу модели артикуляции была положена структура мышц, управляющих артикуляторами *(Рис. 1)*.

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-сжимателя 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 рисунке 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 рисунке1 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 *transversal*. На рис. 1 справа внизу показан поперечный разрез языка примерно в области *genioglossus medialis*, в котором можно видеть мышцу *transversalis*, изгибающую язык в поперечном направлении, создавая как выпуклость, так и впадину. Изгиб описывается «поперечной собственной функцией» в виде половины синусоиды. Этот параметр деформирует поверхность передней трети языка во фронтальной плоскости.

В итоге в экспериментах по решению обратных задач должно использоваться от 16 до 18 артикуляторных параметров. На основе этих параметров модель артикуляции вычисляет форму речевого тракта в средне-сагиттальной плоскости. По расстояниям между подвижными и неподвижными поверхностями речевого тракта вычисляется площадь поперечного сечения речевого тракта S(x,*t*), которая используется решения спектральной задачи для волнового уравнения (1).

4. Моторная теория и теория внутренней модели

Механизмы замыкания обратной связи в системе управления артикуляцией остаются мало изученными, так же, как и механизмы обучения новому языку и компенсации речевой патологии. Гипотеза внутренней модели в системе управления позволяет объяснить наблюдаемые явления. Предполагается, что внутренняя модель является частично врожденным свойством. Она настраивается в период детского лепета, и устанавливает зависимость между нейромоторными командами, сигналами обратной связи от мышечных веретен и проприоцепторов, а также акустическими параметрами сгенерированной речи. В процессе речеобразования, сигналы от рецепторов отображаются в пространство управлений, замыкая таким образом обратную связь. Это отображение выполняется посредством решения так называемой обратной задачи. Результаты исследования свойств речевых обратных задач дают объяснение явлениям компенсации речевой патологии или искусственного возмущения артикуляции или восприятия.

Предполагаемая способность отображения пространства акустических параметров, вычисляемых системой слухового анализа, в пространство управлений может быть распространена на восприятие речи других дикторов, в частности, и на обучение новому языку. Таким образом, гипотеза о моторной компоненте восприятия речи получает поддержку от свойств процессов речеобразования.

Противоречие между наблюдаемым разнообразием акустических параметров и кажущейся устойчивостью восприятия фонетических элементов речи вызвало поиск таких способов обработки речевого сигнала, которые обеспечивали бы меньшую изменчивость описания элементов речи. Эти поиски привели к формулировке различных вариантов моторной теории восприятия речи, предполагающих анализ моторной компоненты речи по речевому сигналу [3–7]. Разные авторы исходили из разных предпосылок, и формировали теории о том, в какой форме моторная компонента принимает участие в восприятии фонетических элементов речи. В основном, представления об участии моторной компоненты опирались на способность человека обучаться речи, слушая речь других людей, хотя бы и с наблюдением за речевой мимикой. Определенную роль сыграло и явление так называемой внутренней речи, т.е. наблюдающееся иногда проговаривание «про себя» читаемого текста. Впоследствии аргументы в пользу восприятия речи на моторном уровне подкрепились установлением того факта, что звуки, воспринимаемые как некая фонетическая единица, обладают сильно различающимися акустическими характеристиками. Вместе с тем, фактически представления о роли моторной компоненты в восприятии речи были гипотезой, а не теорией, поскольку до последнего времени не было прямых доказательств этого явления, и из этих представлений невозможно было сделать прогнозируемые выводы, которые могли быть проверены экспериментально.

Против этой гипотезы было высказано немало возражений, в основном, сводящихся к тому, что ухо получает только акустическую информацию, а больше ничего и не нужно для восприятия речи. На этом основании делается вывод либо о том, что нет необходимости в привлечении моторной компоненты для описания фонетических сегментов [8, 9, 10], либо, что роль моторной компоненты вторична по отношению к акустическим признакам [11,12]. Однако и критика моторной гипотезы также основывалась на умозрительных соображениях, и решающих экспериментальных фактов так и не было представлено.

Были предприняты попытки экспериментального определения того, что же является целью процесса речеобразования — акустический образ или артикуляторные параметры. С этой целью разрабатывались различные методики возмущения артикуляции и восприятия. Исследовались эффекты статического возмущения типа байт-блока, т.е. фиксации положения нижней челюсти [13–16], блока губ с помощью трубочки [17] и искусственного неба [18,19]. Во многих случаях наблюдалась перестройка положений артикуляторов с целью сохранения акустического образа, характерного для произносимого звука или звукосочетания. Аналогичные эффекты были обнаружены и при динамическом возмущении артикуляции. Так, в [20] исследовалось влияние неожиданного возмущения движений губ при артикуляции первого согласного в звукосочетании /i'*pip*/ на сведение и разведение голосовых складок. Было найдено, что при задержке сближения губ разведение складок задерживается, а длительность сведения складок увеличивается. Это наблюдение может быть интерпретировано, как стремление системы управления артикуляцией сохранить акустические характеристики глухого взрыва.

Таким образом, эксперименты с возмущением артикуляции указывали скорее на доминирующую роль акустических параметров, и моторные гипотезы не получили экспериментальной поддержки.

Еще одно возражение против использования моторной компоненты в восприятии можно выдвинуть на том основании, что вычисление этой компоненты есть не что иное, как решение обратной задачи. Хорошо известно, что все речевые обратные задачи являются некорректными в силу неоднозначности отображения пространства акустических параметров в пространство артикуляторных параметров. Поэтому неоднозначность и неустойчивость вычисления артикуляции по акустике кажется непреодолимым препятствием. К тому же решение обратной задачи представляется настолько сложным, что, «по соображениям экономии», кажется нецелесообразным даже в том случае, если удастся найти подходящее решение.

Вместе с тем, имеется и другая группа наблюдений и экспериментов. После удаления гортани по поводу рака, длина и форма речевого тракта изменяются. Тем не менее, в работе [21] были описаны случаи быстрого восстановления акустических характеристик гласных пациентов через две недели после удаления гортани. Это может свидетельствовать о коррекции артикуляции в соответствии с восприятием собственной речи. Через два года формантные частоты гласных этих пациентов были даже ближе к фонетической норме, чем перед операцией. Были также проведены эксперименты с возмущением акустических параметров речевого сигнала — основного тона [22,23] или формантных частот [24, 25, 26]. В этих экспериментах наблюдалась текущая адаптация формы речевого тракта с целью компенсации акустических возмущений. Эти данные свидетельствуют о том, что система управления артикуляцией каким-то образом пересчитывает входные акустические параметры собственной речи в нейромоторные команды.

Рассмотрим эту проблему с точки зрения кодовой структуры речи и примем во внимание погрешности вычисления акустических параметров, а также разнообразие произношения разными дикторами. Речевой поток представляет собой иерархический код с исправлением ошибок. Некоторые элементы этого кода, такие как признаки фрикативных, гласных, смычек и назальных, сравнительно легко определяются на акустическом уровне, хотя их автоматическое распознавание не идеально (см., например, [27). Исследования в области автоматического распознавание речи показали, что место артикуляции взрывных согласных находится по акустическим параметрам наименее надежно. Оценка потенциальной надежности распознавания слов была выполнена в [28, 29] на основе теоремы о кодировании и результатов восприятия бессмысленных звукосочетаний при различных отношениях сигнал/шум. Было установлено, что, при надлежащей лексической избыточности и низком уровне шумов акустические признаки обеспечивают достаточно высокую надежность распознавания слов. При повышении уровня шумов необходимо использовать информацию о месте артикуляции. Если существует возможность определения формы речевого тракта по речевому сигналу, то и нахождение места артикуляции не должно представлять трудности.

Результаты этого анализа находят подтверждение в экспериментах по измерению активности коры головного мозга с использованием функциональной магнитно-резонансной томографии. Было установлено, что при восприятии речи в условиях шумов возникает активность в моторной зоне коры головного мозга, тогда как при хороших акустических условиях активизируется только область слуховой коры [30, 31]. Эти наблюдения служат непосредственным доказательством основного положения моторной гипотезы о том, что в распознавании речи человеком могут принимать участие и моторные компоненты. Аудио-визуальные эффекты также могут свидетельствовать о связи акустического и артикуляторного анализа. 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 [32, 33]. Дальнейшие исследования активности слуховой и моторной зоны коры могут принести ценную информацию о свойствах восприятия речи.

Итак, к настоящему моменту накопились теоретические и экспериментальные результаты, свидетельствующие как о необходимости, так и реальности анализа моторной компоненты при восприятии речи человеком. Однако, механизмы решения обратной задачи — «от акустических параметров к артикуляторным», оставались совершенно неясными. Некоторое представление о таких механизмах можно получить из анализа процессов речеобразования.

Компенсация естественных и искусственных нарушений процесса речеобразования или восприятия является характерным свойством системы управления речеобразованием. Неврологам и логопедам было давно известно, что при парезе или параличе отдельных лицевых или внутриротовых мышц речь может не пострадать. Например, при парезе мышц, управляющих движениями нижней челюсти, артикуляция губных звуков осуществляется за счет большей амплитуды движений губ. Начиная носить зубные протезы с искусственным твердым небом, в ряде случаев люди сохраняли разборчивость своей речи. Иногда больные с удаленной гортанью не только полностью восстанавливали в своей речи различие между звонкими и глухими согласными, но и правильную фразовую интонацию [21], и даже могли петь. Имеются сведения о том, что замена удаленного языка пластиковым протезом позволила больному сохранить сравнительно разборчивую речь.

Реакция артикуляторов на неожиданное механическое возмущение движений губ и нижней челюсти исследовалась в [20, 34–38]. Исследовался также отклик системы управления артикуляцией на изменение формы твердого неба [39] или внезапную электрическую стимуляцию мышц, управляющих артикуляторами [40]. Эксперименты такого типа показывают, что компенсация возмущений может происходить достаточно быстро, с задержкой 10–40 мс, что не оставляет времени для пробных артикуляторных движений, и сопоставимо с общей задержкой прохождения сигнала от периферии до центральной нервной системы и обратно. Известно, что время от сигнала обратной связи мышечного веретена до реакции мышц языка не более 20 мс, а для мышц нижней челюсти — около 15 мс [41]. Это означает, что на артикуляторном уровне сигналы компенсации вычисляются практически мгновенно. При восприятии собственной речи, искаженной частотными или временными преобразованиями также требуется некоторое время для пересчета сигналов рассогласования между ожидаемыми и реально воспринятыми акустическими параметрами. То, что акустическая обратная связь присутствует при оценке собственной речи, следует хотя бы из эффекта возникновения заикания при восприятии задержанной собственной речи [42].

Эти эксперименты показали несостоятельность механизма непосредственной обратной связи, т.е. подачи сигнала от мышечных рецепторов на вход системы управления артикуляцией. Вообще говоря, этого следовало ожидать, поскольку размерность пространства и физическая природа выходного сигнала (смещение артикуляторных органов или акустические параметры речевого сигнала) не совпадают с размерностью пространства и физической природой команд управления сокращением двигательных единиц мышц. Очевидно, что замыкание обратной связи происходит путем пересчета выходных сигналов во входные с помощью некоторого модуля. Необходимость существования такого модуля и его функции описываются гипотезой внутренней модели.

Аналогичная проблема согласования размерности сигналов обратной связи от проприоцепторов и мышечных веретен с размерностью нейромоторных сигналов возникает и при исследовании механизма управления движениями человека. Решение этой проблемы ищут путем введения понятия «схемы тела» или «внутренней модели тела» [43, 44]. Предполагается, что внутренняя модель располагает сведениями о механических свойствах управляемых органов. Она использует их для текущего контроля путем пересчета сигналов от механорецепторов. Эта модель также может порождать сигналы предсказания (feedforward) с целью компенсации нарушений процесса управления. В поддержку мнения о врожденном механизме формирования внутренней модели тела можно привести сведения о фантомах врожденно-отсутствующих конечностей [45, 46]. Следует заметить, что модель управляемого объекта, включенная в систему обратной связи, является важным элементом в теории автоматического управления. Возможно, что идеи из технической области были восприняты в среде исследователей движений человека.

Результаты экспериментов с возмущением артикуляторных движений, упомянутые выше, привели к формулировке гипотезы внутренней модели в управлении артикуляцией. Рассматривались различные нейрофизиологические аспекты этой модели [41, 35]. Математические аспекты гипотезы впервые рассматривались в [28].

Наблюдения за адаптацией оперированных больных к потере голосового источника [21] и компенсацией артикуляции в экспериментах с искусственным возмущением или препятствием движению артикуляторов свидетельствуют о том, что система управления артикуляцией располагает избыточностью на многих уровнях. Одно и то же усилие, развиваемое мышцей, может порождаться активизацией различных двигательных единиц. Одно и то же смещение артикулятора может достигаться за счет разного сочетания сокращения мышц, управляющих его движениями. Наименьшая площадь поперечного сечения речевого тракта в определенном месте может достигаться при разном смещении артикуляторов. Например, одинаковое расстояние между губами достигается различными сочетаниями смещения верхней и нижней губы, а также нижней челюсти. Наконец, одно и то же сочетание резонансных частот речевого тракта может быть получено при разных его формах. Такая избыточность обеспечивает надежность процесса речеобразования и его устойчивость к различного рода патологиям и возмущениям.

Анализ явлений компенсации речевой патологии и внешних возмущений показал, что внутренняя модель может генерировать сигналы обратной связи и контролировать качество речи в текущем времени только в том случае, если она умеет решать так называемую обратную задачу — от выходного сигнала к управлению. В частном случае обратная задача решается, когда входом служат сигналы от механорецепторов. Это позволяет объяснить компенсацию байт-блока. Обратная задача относительно формы речевого тракта или команд управления может решаться и тогда, когда входом являются акустические параметры речевого сигнала, а выходом — форма речевого тракта, артикуляторные параметры или команды управления. Это позволяет объяснить эффекты компенсации возмущения акустических параметров речевого сигнала. Из наблюдений [21] также следует, что и сама внутренняя модель может перестраиваться в новых условиях речеобразования.

Способность внутренней модели к контролю процесса речеобразования можно попытаться распространить и на восприятие речи другого человека, предполагая, что обучение речи или языку «на слух» происходит путем пересчета акустических параметров чужой речи в артикуляторные параметры собственного речевого тракта и установления соответствия с фонетическими элементами языка. Так усматривается прямая аналогия с основными положениями гипотезы о моторной компоненты в восприятия речи. Может пока-

заться, что это является попыткой объяснить одну гипотезу — о моторной компоненте восприятия речи, другой гипотезой — о существовании внутренней модели. Однако предположение об участии внутренней модели в процессах управления артикуляцией в последнее время приобретает характер теории, поскольку позволяет осмыслить факты, не поддающиеся истолкованию другим образом.

Если способность к формированию внутренней модели является врожденным свойством, то настройка ее параметров может происходить в период детского лепета путем вариации всех возможных артикуляторных управлений и запоминанием соответствующих акустических параметров. По мере роста речевого тракта параметры внутренней модели уточняются, но и предыдущие могут сохраняться, что облегчит решение обратной задачи для речи других людей. Таким образом, восприятие чужой речи может, по крайней мере, частично, выполняться в терминах собственной внутренней модели.

Избыточность управления речеобразованием означает, что при попытке определения формы речевого тракта, артикуляторных параметров или команд управления не только по акустическим параметрам речевого сигнала, но и по сигналам обратной связи от механорецепторов, принципиально отсутствует однозначное решение. В математике такие задачи называются некорректными. Кинематическая неоднозначность, казалось бы, делает бессмысленной постановку обратной задачи для речевого тракта и в значительной степени обесценивает как концепцию внутренней модели, так и гипотезу о моторной компоненте в восприятии.

Тем не менее, речевые обратные задачи относительно формы речевого тракта, положения артикуляторов и команд управления могут быть решены с достаточной точностью.

5. Речевые обратные задачи

В силу кинематической неоднозначности все обратные задачи для речи являются некорректными по Адамару, т.е. формально для них не гарантируется однозначное и устойчивое решение волнового уравнения относительно площади поперечного сечения речевого тракта и, тем более, относительно артикуляторных параметров. Однако, вариационный метод и регуляризация по Тихонову [47], в совокупности с сильными ограничениями на значения и динамику артикуляторных параметров, позволяют получить устойчивые и достаточно точные решения речевых обратных задач. Вариационный метод требует использования математических моделей процессов речеобразования, и это совпадает с гипотезой существования таких моделей в системе управления артикуляцией. Эта модель задается в виде

$$A(x)=u,$$

где x — артикуляторные параметры, u — акустические параметры.

В методе Тихонова ищется приближенное решение обратной задачи путем минимизации функционала

$$M(z) = \alpha\Omega(z) + \rho^2(A_h z, u_\delta), \quad z \in Z \qquad (2)$$

где A_h — оператор приближенной (с точностью h) математической модели, связывающей входные параметры инвертируемого процесса z и выходные параметры u_d, измеренные с погрешностью δ. W(z) есть критерий оптимальности, a=a(h,δ) — параметр регуляризации. Величина $\rho(A_h z, u_\delta) = \|A_h z - u_\delta\|$ есть невязка между измеренными и вычисленными параметрами, а Z — данное множество ограничений. В нашем случае h и δ — погрешность в описании модели речеобразования и ошибки измерения акустических параметров.

Процесс минимизации состоит в поиске условного экстремума при наличии ограничений на значения артикуляторных и акустических параметров.

Критерий минимума работы артикуляторов оказался эффективным при решении обратных задач для стационарных сегментов гласных или фрикативных звуков. На *рис. 2* показаны профили речевого тракта в средне-сагиттальной плоскости, измеренные с помощью рентгенографии, и вычисленные формы тракта.

При решении динамических задач необходимо использовать составной критерий $\Omega = \Omega_W + \Omega_T$, где

$$\Omega_W = \frac{1}{2T}\sum_k \int_t^{t+T} c_k (x_k - x_k^{(o)})^2 d\tau$$

$$\Omega_T = \frac{1}{2T}\sum_k \int_t^{t+T} (m_k x_k'')^2 d\tau$$

Здесь C_k — коэффициент упругого сопротивления движению артикулятора, m_k — масса артикулятора, $x_k^{(0)}$ — значение артикуляторного параметра в нейтральном состоянии. Эти критерии интерпретируются соответственно как средняя за время T суммарная работа упругих сил (Ω_W) и средний квадрат полной силы, приложенной к артикуляторам (Ω_T) [48].

Решение динамических обратных задач в ряде случаев также оказывается вполне удовлетворительным. Ошибка аппроксимации движений некоторых точек внутри речевого тракта, измеренных с помощью микролучевого рентгеноскопа, и акус-

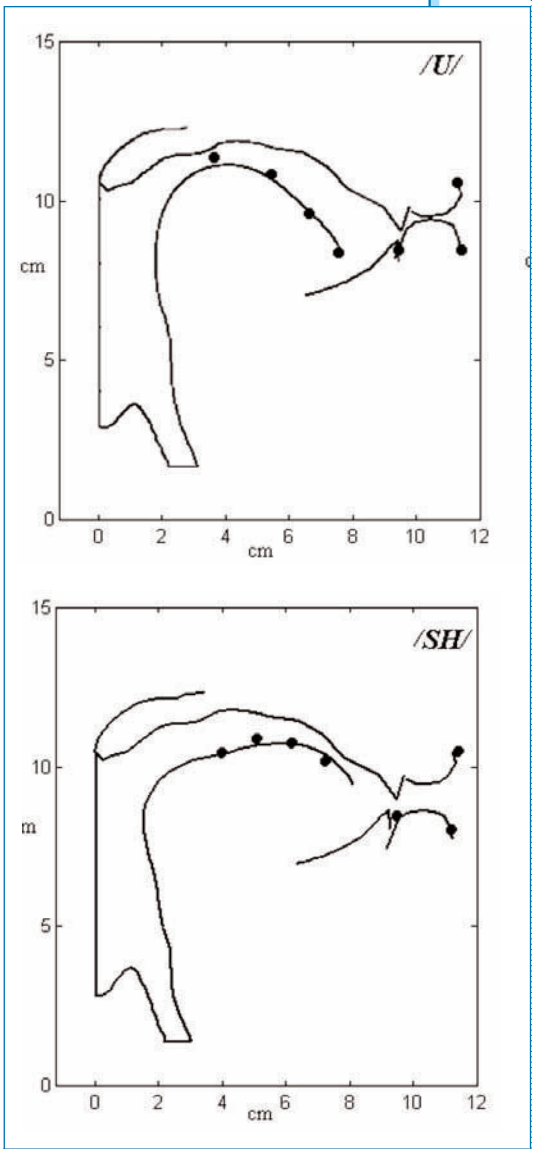

Рис. 2. Сравнение решений обратной задачи для гласных и рентгенографических измерений.

тических параметров находятся в пределах погрешности измерений. Гласные, сгенерированные артикуляторным синтезатором по результатам решения обратной задачи, субъективно оказываются весьма похожими на оригинальные звуки [49]. Решение динамической обратной задачи для слогов, содержащих фрикативные, также демонстрирует высокую точность восстановления артикуляторных и акустических параметров и похожее звучание. На *рис. 3* показаны сонограммы оригинальных слогов */аша/*, */аса/* и сонограммы слогов, ресинтезированных по решению обратной задачи.

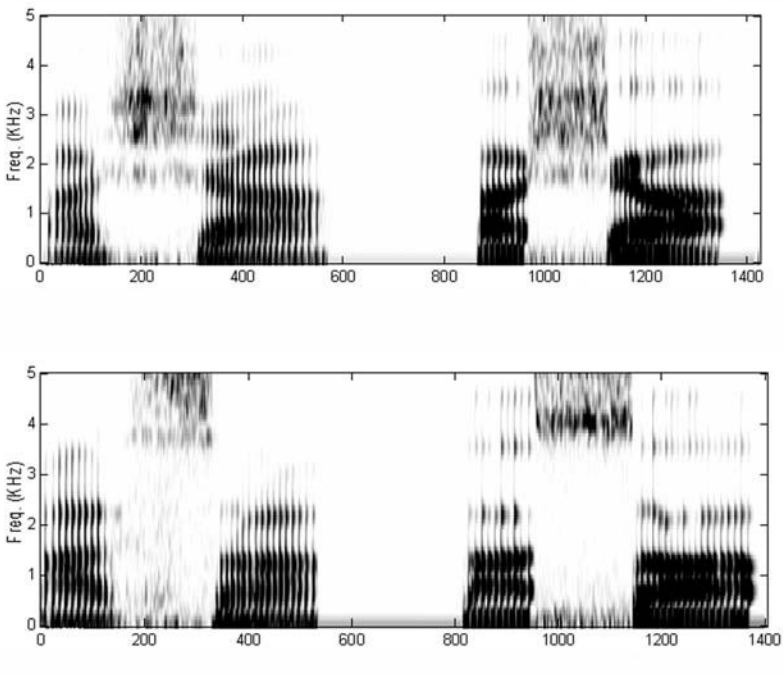

Рис. 3. Сонограммы оригинальных (слева) и синтезированных (справа) по решению обратной задачи слогов /аса/ (вверху) и /аша/ (внизу).

Задача минимизации функционала (2) оказывается многоэкстремальной, и нет гарантий нахождения глобального минимума. Для нахождения локального минимума, обеспечивающего приемлемую точность решения, необходимо повторять процесс оптимизации определенное количество раз, начиная с разных начальных условий. Начальные значения выбираются не произвольно, а из так называемой кодовой книги, в которой каждому вектору акустических параметров соответствует некоторое множество векторов артикуляторных параметров. Первоначально, при исследовании речевых обратных задач кодовая книга формировалась с использованием артикуляторного синтезатора [50]. Однако при этом необходимо было либо произвольно задавать геометрические размеры речевого тракта, либо подставлять значения, измеренные на конкретном дикторе. Попытки применения фиксированных анатомических параметров к разным дикторам показали необходимость создания кодовой книги, в которой, помимо артикуляторных параметров, присутствуют и анатомические параметры для разных дикторов [51]. Таким образом, поиск методов решения обратной задачи для речевого тракта привел к традиционному для распознавания речи подходу, т.е. к необходимости обучения кодовой книги на представительном множестве дикторов.

Были также найдены условия, при которых погрешность полученного решения не превышает удвоенную погрешность потенциально достижимого решения, хотя эта потенциальная точность остается неизвестной. Следовательно, вопрос о возможности решения обратной задачи сводится к оценке погрешности решения. Если эта погрешность, например, находится в пределах погрешности измерения, то обратную задачу можно считать решенной.

Как видно, в методе регуляризации важную роль играет наличие математической модели процесса, который должен быть инвертирован. Именно эта модель обеспечивает, наряду с другими приемами, единственность и устойчивость решения обратной задачи. Таким образом, объясняется возможный механизм действия внутренней модели в системе управления речеобразованием как для контроля собственной речи, так и для восприятия речи других людей.

При решении динамической обратной задачи, т.е. задачи относительно управлений, необходимо создать некоторую модель управлений. Используемая *динамическая модель* связывает переменный вектор артикуляторных параметров $x(t)=(x_1(t),\ldots,x_{n(t)})$ с вектором $u(t)=(u_1(t),\ldots,u_{n(t)})$ управляющих воздействий посредством системы обыкновенных дифференциальных уравнений

$$x_i'' + 2g_i x_i' + \omega_i^2 x_i = u_i(t), \quad i = 1,\ldots,n \qquad (3)$$

Параметры g_i и ω_i системы (3) характеризуют динамические свойства *i*-го артикулятора. Координата u_i вектора управления интерпретируется как ускорение, $u_i = G_i / m_i$, создаваемое силой G_i, которая развивается мышцами, связанными с *i*-м артикулятором массы m_i. Уравнение (3) описывает динамику артикуляторов только приблизительно, поскольку и потери и упругое сопротивление зависят от приложенной силы. Однако, при качественных исследованиях динамических свойств артикуляторов уравнение (3) приемлемо.

Поиск модели управлений опирался на данные о том, что команды на исполнение новой программы движений человека не могут поступить раньше, чем через некоторое время после активизации предыдущей программы [52]. Это означает, что команды могут быть разрывными во времени, или «кусочными». В [53] исследовался функциональный класс управлений в виде разрывных во времени полиномов вплоть до третьей степени:

$$u_i(t) = a_{0i} + a_{1ij}(t - t_j) + a_{2ij}(t - t_j)^2 + a_{3ij}(t - t_j)^3, \quad t \in [t_j, t_{j+1}]. \qquad (4)$$

Экспериментальными данными служили измерения движений небной занавески, нижней челюсти, кончика языка и нижней губы, выполненные на микролучевом рентгеноскопе совместно с регистрацией ЭМГ мышц *levator palatini, longitudinalis superior, longitudinalis inferior, masseter.* Было установлено, что полином нулевого порядка, т.е. ступенчатое возбуждение приводит к значительному перерегулированию в переходных процессах *(рис. 4).* Наименьшую ошибку в аппроксимации движений артикуляторов обеспечивает полином первого порядка *(рис. 5).* При этом вычисленные величины управляющих сил находятся в физиологически правдоподобных пределах, а форма вычисленных управлений коррелирована с электрической активностью мышцы в тех случаях, когда она является единственной для данного движения артикулятора.

Кодовая книга

Математически, минимизация функционала (2) рассматривается как задача поиска условного экстремума критерия оптимальности при различного рода ограничениях. Речевая обратная задача является многоэкстремальной вследствие неоднозначности отображения пространства акустических параметров в пространство артикуляций. Стандартный

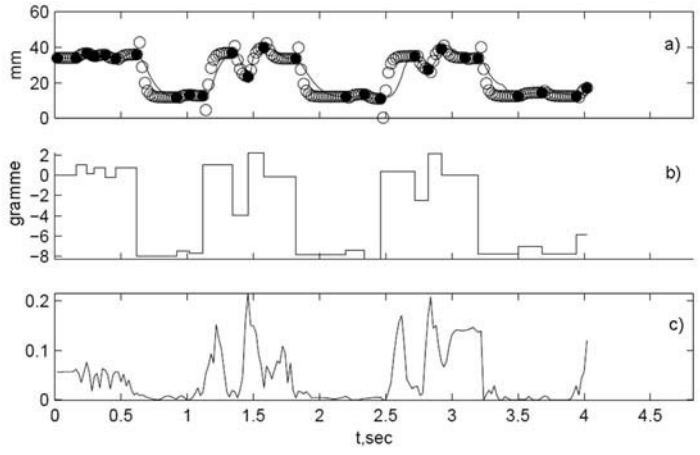

Рис. 4. Ступенчатое управление движений нёбной занавески

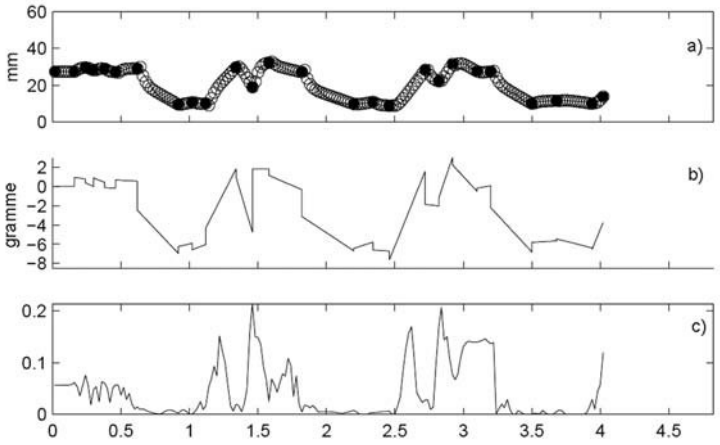

Рис. 5. Линейное управление движений нёбной занавески

подход к решению таких задач состоит в использовании некоторого множества начальных приближений и последующем сравнении его результатов. Это позволяет если не найти глобальный экстремум, то хотя бы выбрать решение с наиболее подходящими характеристиками.

Выбор начальных приближений сам по себе является непростой задачей, которая может быть облегчена путём создания так называемой кодовой книги, в которой содержатся векторы артикуляторных параметров и соответствующие им векторы акустических параметров. Поскольку кодовая книга должна содержать конечное число ячеек, то и артикуляторные, и акустические параметры квантуются с некоторой точностью. Если имеется достаточно обоснованная модель речеобразования, то формирование кодовой книги происходит путём многократного решения относительно простой прямой задачи, т.е. вычисления акустических характеристик для артикуляторных параметров, перебираемых с некоторым шагом. Метод кодовой книги для решения речевых обратных задач был впервые предложен в [50] и затем использовался в [54–58].

В формировании кодовой книги значительную роль играют анатомические размеры речевого тракта, для которого она составляется. Кодовая книга для решения обратных задач внутренней моделью в системе управления артикуляцией конкретного диктора может быть построена сравнительно легко. Для этого достаточно запомнить команды управления и соответствующие им акустические параметры сгенерированного речевого сигнала. Если не требуется особенно высокая точность решения обратной задачи, то кодовая книга позволяет избежать трудоёмкого процесса оптимизации путём простого выбора команды, соответствующей своему услышанному речевому сигналу, а затем сравнить его с поданным сигналом, замыкая обратную связь. Локальная линеа-

ризация модели речеобразования для каждой ячейки кодовой книги позволяет существенно ускорить процесс решения обратной задачи [59]. Такая линеаризация дает более точное решение обратной задачи.

Таким образом, кодовая книга может рассматриваться не только как формальный инструмент для хранения начальных приближений, но и как механизм решения речевых обратных задач, который может иметь физиологическую основу.

Построение кодовой книги для определения моторной компоненты при восприятии речи других людей, очевидно, является более трудной задачей, но и она может быть решена. Во-первых, собственная кодовая книга, созданная для внутренней модели, может использоваться для восприятия речи людей с похожей анатомией. Во-вторых, в процессе роста собственного речевого аппарата могут быть созданы кодовые книги, соответствующие разным его размерам. Это способствует восприятию речи детей и людей с меньшими размерами тракта. Наконец, 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 [60]. Наблюдение за лицом диктора влияет на восприятие речи, и в случае противоречия между видимой артикуляцией и услышанным звукосочетанием возникают разнообразные эффекты восприятия [61].Информативность наблюдаемых проявлений артикуляции позволяет общаться глухонемым.

Таким образом, кодовая книга, созданная для обеспечения управления артикуляцией, может быть дополнена в процессах обучения пониманию речи других людей. То, что этот механизм не исчезает после периода становления речи у детей, подтверждается способностью к усвоению иностранных языков и приспособлению к пониманию речи людей с особенностями произношения.

В описываемых экспериментах использовалась кодовая книга, созданная на основе измерений траекторий координат [58] нескольких точек на языке, губах, верхних и нижних зубах с помощью микролучевого рентгеноскопа [62]. Синхронное измерение акустических параметров позволяет поставить задачу формирования кодовой книги для реальных дикторов как специфическую обратную задачу. Решение этой задачи значительно легче, чем решение задачи, когда входными данными служат только акустические параметры, особенно, если при этом доступны и измерения формы твёрдого нёба в среднесагиттальной плоскости, формы челюсти в латеральной плоскости, расстояния от передних зубов до задней поверхности тракта и положения гортани.

Статические и динамические обратные задачи

В работах [49, 51, 57, 59] исследовались обратные задачи для гласных звуков. В работах [63, 64] решались обратные задачи для фрикативных. В целом, точность вычисления координат точек измерения на квазистационарных участках гласных составила около 2.8%, что находится в диапазоне погрешности измерения. Разница между измеренными и вычисленными формантными частотами в среднем была 7.7% для F_1, 3.8% для F_2, и 2.6% — для F_3. Точность воспроизведения координат точек измерения для фрикативных также находилась в пределах погрешности измерения — около 3%. Погрешность определения характерных частот спектров была около 8.5%. На *рис. 6* показаны сонограммы слогов исходной речи и речи, синтезированной по результатам решения обратной задачи.

Эти результаты указывают на принципиальную возможность достаточно точного решения речевых обратных задач.

Прежде чем пытаться распространить концепцию внутренней модели на процессы восприятия, следовало бы убедиться в том, что внутренняя модель действительно может решать обратные задачи типа «акустические параметры — форма речевого тракта», «форма речевого тракта — артикуляторные параметры», «артикуляторные параметры — управления» в случае, когда анатомические размеры и динамические параметры артикуляторов известны. С этой целью, во-первых, нужно решить эти задачи с приемлемой точностью, пользуясь только физиологически правдоподобными параметрами. Во-вторых, степень доказательности концепции внутренней модели существенно возросла бы, если бы удалось воспроизвести некоторые явления и эффекты, не находящие объяснения в рамках других подходов к описанию свойств системы управления речеобразованием. Компьютерное моделирование эффектов ограничения на движение нижней челюсти (bite-block) и реорганизации управлений было описано в [65, 66].

Bite-block

В известных экспериментах [13], препятствие к подъёму нижней челюсти при артикуляции гласных (bite-block) приводило к компенсационным движениям губ и языка так, что акустические характеристики гласных незначительно отличались от исходных. Способность критериев оптимальности, используемых для решения обратных задач, к воспроизведению эффекта компенсации bite-block на материале кинорентгенограмм исследовалась в [65, 66]. Из слитных фраз вырезались участки длиной в несколько слогов, и для них измерялись расстояния между губами, передними зубами, кончиком языка и твёрдым нёбом, средней части языка и нёбной занавеской. Точность аппроксимации этих траекторий при решении обратной задачи относительно команд управления была в пределах точности измерений. Симуляция bite-block осуществлялась путём фиксации расстояния между передними зубами (1 см) с попыткой решить обратную задачу, требуя достижения смычки на губах, передней части языка или в области мягкого нёба. Результаты решения задачи для звукосочетания /*паникапа*/ показаны на *рис. 6*. Как видно, несмотря на фиксацию нижней челюсти, все три смычки — губная, переднеязычная и заднеязычная были сформированы, и даже траектории исходных движений и решений обратной задачи оказались очень близки.

Это объясняет эффект мгновенной компенсации (уже на первом импульсе голосового источника), обнаруженной в экспериментах [13]. По-видимому, на этапе движения артикуляторов от нейтрального состояния к состоянию, характерному для какого-то гласного, ещё до включения голосового источника система управления артикуляцией обнаруживает фиксацию нижней челюсти и перестраивает партитуру команд согласно заданному критерию оптимальности.

Ускорение артикуляции

ЭМГ измерения потенциалов мышц-артикуляторов показывают, что изменение темпа артикуляции приводит к перераспределению активности

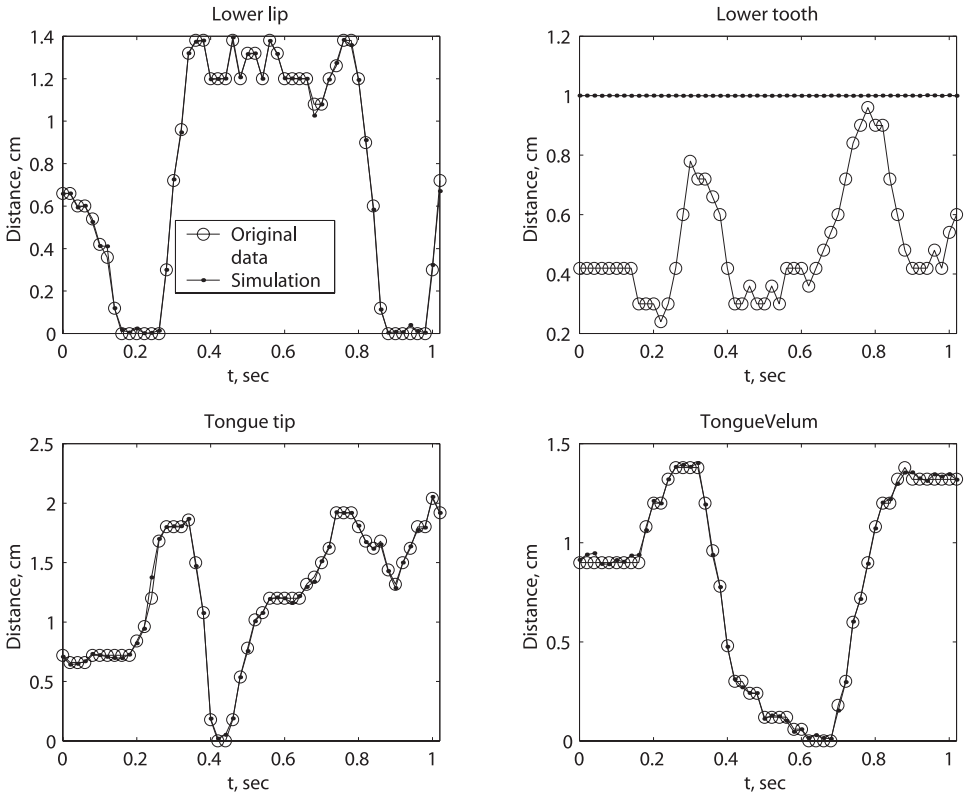

Рис. 6. Симуляция bite-block для звукосочетания /паникапа/: a) расстояние между губами; b) расстояние между передними зубами; c) расстояние между кончиком языка и твёрдым нёбом; d) расстояние между языком и нёбной занавеской

мышц и движений артикуляторов [28, 67, 68]. Очевидно, что статические критерии оптимальности не могут воспроизвести эффект реорганизации управлений, поскольку используют информацию только об установившемся состоянии артикуляторов. Можно ожидать, что динамические критерии продемонстрируют реорганизацию управлений.

В [66] ускорение артикуляции симулировалось прореживанием отсчётов рентгенографических измерений при фиксированном темпе артикуляции. Эффект реорганизации управлений наблюдался во всех экспериментах. На *рис. 7* показаны управления для исходного дифтонга /*ai*/ из базы данных, полученной на микролучевом рентгеноскопе [16], и управления для вдвое «ускоренной» артикуляции этого дифтонга. Входными параметрами для решения обратной задачи здесь служили измерения трёх формантных частот и траектории нескольких маркеров внутри речевого тракта. Как видно, форма и фазы команд относительно друг друга значительно изменились, за исключением кончика языка. Отсутствие реальных данных для одного и того же звукосочетания, произнесённого в разных темпах, не позволяет судить о точности воспроизведения эффекта реорганизации управлений, однако сам эффект налицо.

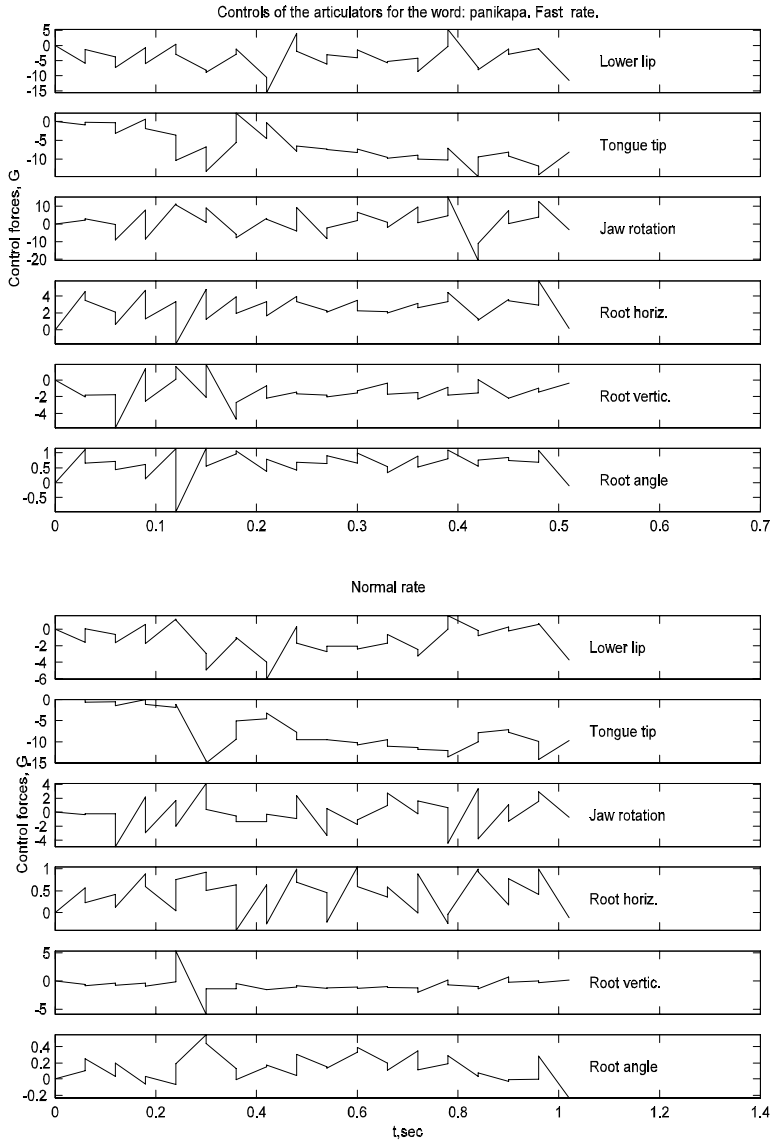

Рис. 7. Симуляция реорганизации управлений при ускорении артикуляции для дифтонга /ai/, вверху — ускоренный темп внизу — нормальный темп: 1) нижняя губа; 2) кончик языка; 3) угол поворота нижней челюсти, 4) горизонтальное смещение корня языка; 5) вертикальное смещение корня языка; 6) угол поворота языка как твёрдого тела. Управление — в граммах силы

6. Модель периферического анализа речевого сигнала

Новый интерес к разработке математических моделей восприятия связан с задачей повышения помехоустойчивости систем автоматического распознавания речи в связи с неспособностью метода скрытых марковских моделей к подавлению помех и искажений. В частности, сообщается о значительном повышении устойчивости к шуму при использовании нелинейных моделей периферического отдела слухового анализатора [69–73].

Имеется ряд хорошо установленных свойств восприятия речи и неречевых стимулов. Сюда относятся эффекты адаптации к акустическим свойствам канала связи, включения и выключения стимула, прямой и обратной временной маскировки, спектрально-временного латерального торможения, логарифмическая шкала частот и амплитуд, а также сведения о существовании детекторов амплитудных и частотных модуляций. Это позволяет создать феноменологическую модель первичного анализа речи, обладающую малой чувствительностью к квазистационарным амплитудно-частотным характеристикам канала связи. В этой модели используются только операции задержки во времени, усреднения по времени или по частоте, а также логарифмирование [74].

Оператор

$$A(\omega,t)=\lg\frac{S(\omega+\Delta\Omega,\theta_1,t\pm\Delta T_1,\tau_1)+C}{S(\omega-\Delta\Omega,\theta_2,t\mp\Delta T_2,\tau_2)+C}, \qquad (5)$$

описывает акустические (неспецифические) детекторы спектрально-временных неоднородностей сигнала и моделирует многие известные свойства слухово-

го восприятия. Здесь S — спектр мощности принятого сигнала, очищенного от аддитивных шумов; $\Delta\Omega$ — сдвиг отсчёта спектра по частоте; ΔT — сдвиг отсчёта спектра по времени; θ_1 и θ_2 — скользящие интервалы сглаживания спектра по частоте; τ_1 и τ_2 — постоянные времени сглаживания спектральных компонент фильтром первого порядка, $C \geq 1$.

На *рис. 8* показаны результаты обработки речевого сигнала для последовательности слов «один, шесть, четыре» оператором (5) с разными параметрами. Апостроф на символах разметки означает мягкие согласные, символ *vh* обозначает аспирацию в конце слова, символ *Th* — аспиративный взрыв, символ *T!* — неаспиративный взрыв, а символ *z#* — начало паузы между словами. Под речевым сигналом изображена сонограмма, вычисленная в шкале мелов и сглаженная по частоте треугольными фильтрами (наклон +25 дБ/Барк и — 10 дБ/Барк). Под сонограммой показаны «детектограммы» — положительные отклики оператора с параметрами $\theta_1 = 0$, $\theta_2 = 0$, $\Delta\Omega = 0$, $\tau_1 = 5$ мс, $\tau_2 = 15$ мс, $\Delta T_1 = 0$,

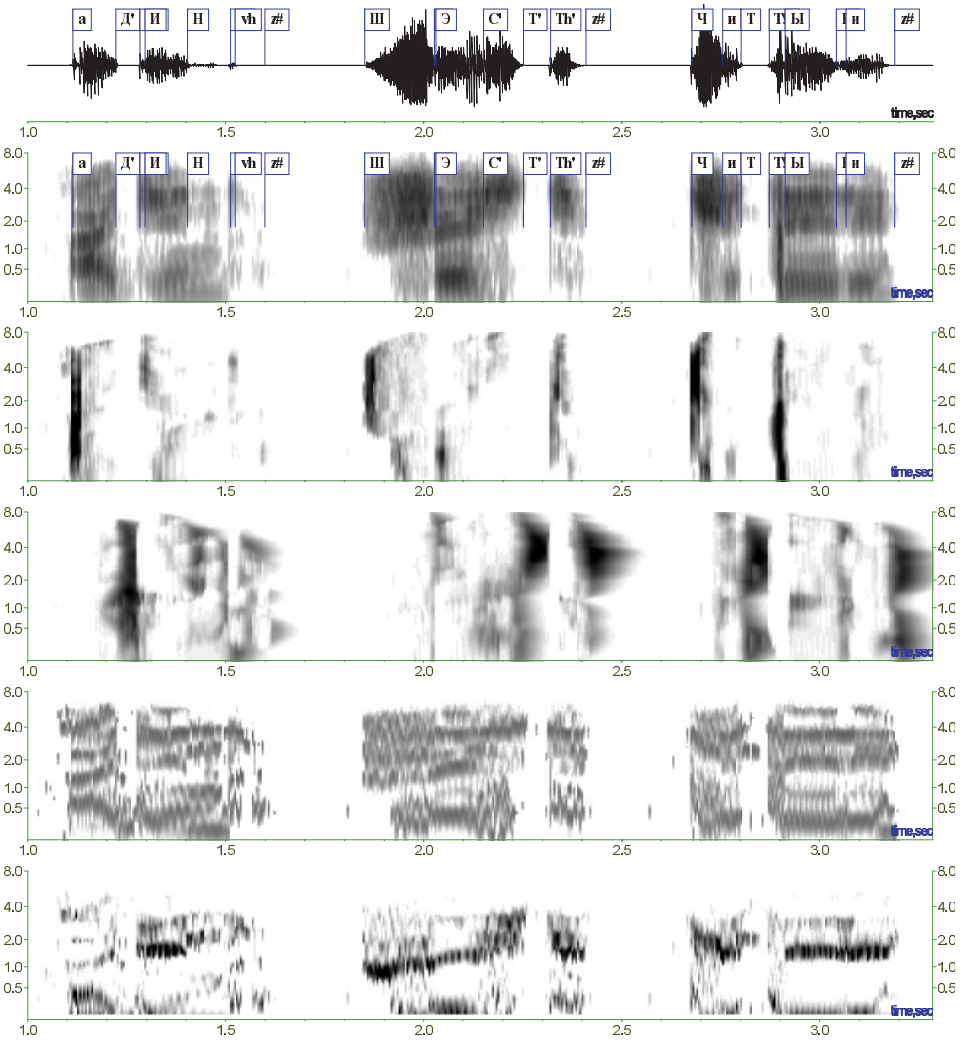

Рис. 8. Слова «один, шесть, четыре». Сверху вниз: осциллограмма речевого сигнала; сонограмма; отклик динамического детектора на возрастание энергии; отклик динамического детектора на спад энергии; сонограмма, нормированная к скользящему по частоте интервалу; конечная разность по частоте (пояснения в тексте)

ΔT_2 = -25 мс, и инвертированная «детектограмма» для отрицательных откликов оператора с теми же параметрами. Ниже расположена сонограмма, нормированная по частоте на скользящих интервалах θ_1=40 мел, θ_2=600 мел, а под ней — положительные конечные разности спектра по частоте, вычисленные на интервале 120 мел.

Разные совокупности параметров в операторе (5) выделяют в динамическом спектре речевого сигнала различные виды состояний и переходных процессов. Поэтому можно представить многослойный поток «детектограмм», ориентированных на выделение различных сегментов речевого сигнала. Поскольку каждая «детектограмма» выглядит как двумерное изображение, один из подходов к их использованию состоит в применении методов обработки изображений. Анализ «детектограмм «представляет собой новое направление в сегментации и распознавании речи. Здесь открывается обширное поле для исследований.

7. Синтез речи

Все современные системы синтеза речи по тексту — формантные, компиляционные, гибридные — страдают нарушением динамики важных параметров речевого сигнала, что, в конце концов, приводит не только к ухудшению натуральности и разборчивости, но и к быстрому утомлению слушателя. Очевидно, что если бы удалось построить математическую модель речеобразования, детально описывающую все процессы — акустику разветвлённой системы с податливыми стенками, механику движений артикуляторов, систему управления артикуляцией, аэродинамику воздушного потока, то следовало бы ожидать, что синтезированный по правильным управлениям речевой сигнал практически не отличался от реального сигнала. Здесь важное словосочетание — «правильные управления». Математические модели, описанные в [28, 75], обеспечивают высокую натуральность отдельных гласных и коротких слогов, синтезированных с помощью подобранных вручную правлений. Однако ясно, что для синтеза речи по произвольному тексту невозможно вручную подобрать управления для всех возможных звукосочетаний. Нужны средства, позволяющие автоматизировать процесс формирования команд на артикуляторные органы.

Эти средства появились только после того, как были разработаны методы решения речевой обратной задачи «от речевого сигнала к командам управления», частично описанные выше. На *рис. 9* приведены сонограммы речевого сигнала для фразы «*The other one is too big*». Вверху показана сонограмма для исходного произнесения, а внизу — сонограмма сигнала, синтезированного после решения обратной задачи относительно команд управления.

Ошибки воспроизведения формант этой фразы были около 3%, а спектра фрикативных звуков — около 8%. Исходная и ресинтезированная фраза на слух практически неотличимы, что свидетельствует о приемлемом качестве решения обратной задачи и адекватности математической модели речеобразования.

Для того чтобы создать высококачественный синтезатор речи тексту, необходимо решить следующие задачи:

1. Вычислить команды управления для всевозможных диад или, что лучше, триад. Для этого нужно проанализировать более 10.000 слогов, но объем памяти не имеет значения.
2. Разработать алгоритм перевода буквенного представления текста в фонетическое.
3. Разработать алгоритм сшивания команд управления на границах слогов.
4. Разработать алгоритм вычисления просодических параметров по тексту.

Последняя задача представляется весьма трудной даже для синтеза нейтральной интонации, не говоря уже о чтении художественного текста.

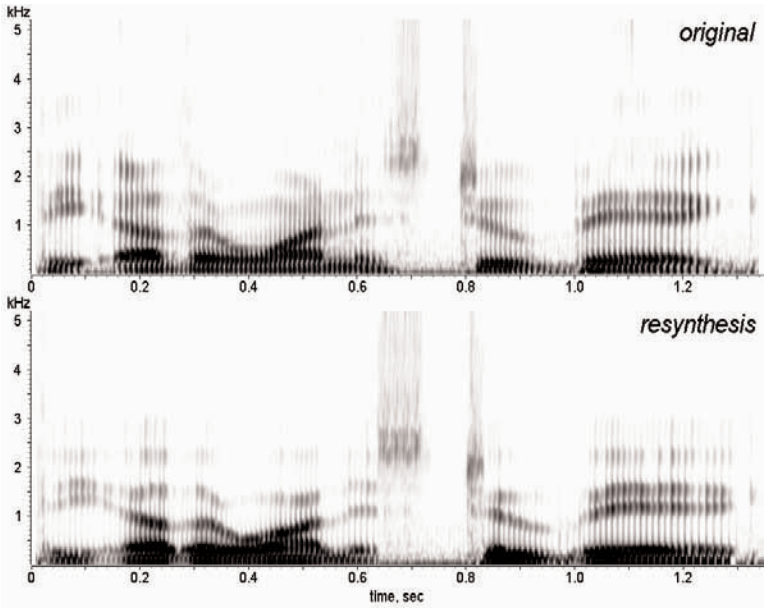

Рис. 9. Сонограммы исходной фразы (вверху) и ресинтезированной по командам, вычисленным путём решения обратной задачи (внизу)

8. Распознавание речи

Современные коммерческие системы автоматического распознавания речи достигли уровня надёжности распознавания слов близкой к 90%, особенно при условии адаптации к диктору. Объём распознаваемого словаря расширился до сотен тысяч словоформ. Основные усилия теперь концентрируются на создании интерфейса «человек — машина», наиболее удобного для пользователя. Складывается впечатление, особенно для внешнего наблюдателя, что принципиальные проблемы распознавания речи уже решены и остались лишь технические и эргономические задачи. Это привело к свёртыванию фундаментальных исследований за рубежом и закрытию исследовательских отделов в крупнейших частных компаниях. Однако появление коммерческих систем распознавания обусловлено не столько прорывом в решении принципиальных проблем, сколько значительным расширением возможностей персональных компьютеров.

Практика применения систем распознавания речи показала, что они неустойчивы к помехам и искажениям канала речевой связи. Типичным является катастрофическое снижение словесной надёжности распознавания до величин порядка 40–60% при появлении относительно слабых шумов, смене типа микрофона или канала связи (см. например, [73, 76]). По некоторым экспертным оценкам, исходная надёжность распознавания слов в диктофоне компании IBM до адаптации к диктору составляет около 70%, что исключает использование этой системы в режиме независимости от диктора. Сравнительно высокая словесная надёжность достигается только при раздельном произнесении слов и только в условиях близких к тем, при которых происходило обучение.

Поэтому высказываются серьёзные сомнения в возможности такого развития метода скрытых марковских моделей (СММ), которое привело бы к окончательному решению зада-

чи распознавания для любых приложений [70, 72]. Эти сомнения вполне обоснованны, поскольку метод СММ, по существу, не является специфическим для речи и не учитывает фундаментальные свойства речи. Этот метод вполне применим к распознаванию любых акустических сигналов, а не только к распознаванию речи. К числу специфических свойств речи, в первую очередь, относится тот факт, что речевой сигнал предназначен для передачи сообщений и содержит в себе код, специально сконструированный для коррекции ошибок, возникших в процессе речеобразования и передачи речевого сигнала по какому-то каналу связи.

Задача распознавания или понимания речи является обратной задачей в том смысле, что по принятому речевому сигналу нужно восстановить фонетический состав или смысл переданного сообщения. Как известно, обратные задачи часто некорректны, т.е. их решение неоднозначно и неустойчиво относительно помех и искажений. Устойчивое решение обратной задачи может быть получено только при условии использования математической модели распознаваемого процесса и определённых ограничений на возможные решения. Это приводит к необходимости разработки моделей процессов речеобразования и восприятия речи, включая модель кодовой структуры речевого сообщения, поскольку для защиты от помех и искажений речь должна обладать свойствами кодов, исправляющих ошибки.

Постановка задачи автоматического распознавания речи зависит от практического приложения. Собственно распознавание речи подразумевает в более широком смысле понимание того, *что* было сказано. Но возможны и другие постановки задачи, например, при распознавании (верификации или идентификации) диктора (*кто* сказал), распознавании состояния диктора (*как* сказал) или распознавании среды, окружающей диктора (*в каких условиях* сказал).

Изменчивость

Трудности в автоматическом распознавании речи связаны с изменчивостью акустического образа, приписываемого одному и тому же речевому элементу, например слову. Существует несколько видов изменчивости, каждая со своими закономерностями. Условно можно различать изменчивость, связанную с внешними условиями, дикторскую изменчивость и контекстную изменчивость. Ниже перечислены наиболее часто встречающиеся виды изменчивости:

- Акустические помехи внешней среды, среди которых наиболее часто встречаются нестационарные помехи в виде речи посторонних дикторов. Борьба с такими помехами, получившими название «coctail party effect», пока не увенчалась успехом.
- Искажение характеристик речевого сигнала в тракте между микрофоном и аналого-цифровым преобразователем. Сюда входят наводки электрических линий и шумы электронных цепей, разные коэффициенты усиления. Особенно велики помехи и замирания, характерные для радиоканалов с аналоговой передачей сигнала.
- Искажения амплитудно-частотных и временных характеристик речевого сигнала в результате реверберации замкнутых помещений. В частности, реверберация приводит к длительному присутствию резонансных колебаний на смычках после гласных звуков.
- Искажение амплитудно-частотных характеристик речевого сигнала, свя-

занное с различием типов микрофонов, расстояния от рта диктора до микрофона и направления микрофона. Близко расположенные микрофоны улучшают отношение «речевой сигнал — акустические шумы среды», однако при этом возникает эффект ближнего акустического поля, при котором амплитудно-частотные характеристики сигнала в низкочастотной области сильно зависят от расстояния до микрофона. Использование головных гарнитур с близко расположенным микрофоном неприемлемо для большинства пользователей.

- Изменчивость амплитудно-частотных характеристик стационарных сегментов речевого сигнала, связанная с различием размеров и формы речевого тракта дикторов.
- Различие в темпе речи дикторов, которая при прочих фиксированных условиях может доходить до 300%. Изменчивость длительности фонетических элементов в зависимости от стиля речи, эмоционального и физического состояния диктора.
- Изменчивость громкости речи диктора и связанная с этим изменчивость амплитудно-частотных характеристик речевого сигнала. В частности, известен так называемый эффект Ломбарда, состоящий в повышении уровня высокочастотных компонент речевого сигнала при непроизвольном повышении громкости при разговоре в присутствии помех.
- Разнообразие динамических характеристик речи, связанное с различием масс артикуляторных органов и особенностями артикуляции дикторов, стилем речи, эмоциональным и физическим состоянием дикторов.
- Изменчивость длительности и акустических характеристик фонетических элементов в зависимости от длительности фразы, положения относительно начала фразы и положения относительно логического ударения во фразе.
- Изменчивость граничных фонетических элементов слов в слитном потоке речи — слияния конечных и начальных фонетических элементов, оглушение, озвончение, назализация и прочие эффекты коартикуляции.

Отсюда, в частности, вытекают требования к формированию базы данных для обучения системы распознавания. Чтобы избежать настройки на фиксированные условия записи речи, база данных должна быть по возможности неоднородной. В ней должны быть представлены разнообразные виды помех и искажений.

Ни один из известных формальных «математических» методов не в состоянии компенсировать все виды изменчивости. Это относится к когда-то популярному методу неоднородной деформации временной оси, скрытым марковским моделям и нейронным сетям. «Физический» подход уделяет большее внимание структуре речевого сигнала и поиску адекватных единиц распознавания. Этот подход в настоящее время в основном представлен системами, построенными на основе экспертных знаний, почерпнутых из опыта чтения сонограмм («видимой речи») (Zue et al., 1990). Эти знания весьма субъективны, и задача борьбы с изменчивостью в явном виде в них не формулируется.

В системах понимания речи и в задачах в ограниченной предметной областью любой метод должен дополняться лингвистическим анализом лексических, грамматических, семантических и прагматических связей в речевом потоке.

Кодовые свойства речи

С точки зрения современной теории кодов, корректирующих ошибки, речь принадлежит к классу нелинейных кодов, поскольку всегда найдётся хотя бы одна пара слов, которая при любом методе их сложения не образует новое осмысленное слово. Декодирование таких кодов возможно только полным перебором всех возможных слов. Такой перебор может быть

реализован с помощью динамического программирования, в частности методом Витерби, или методом последовательного декодирования.

В организации речевого кода просматривается структура, аналогичная каскадным кодам, поскольку коррекция ошибок возможна за счёт использования избыточности на уровне артикуляции (не все последовательности артикуляторных состояний физически реализуемы), признаков фонетических элементов, слогов, слов и фраз. Существуют также уровни семантических и прагматических ограничений. Декодирование речевого сигнала с использованием предсказания, полученного от разных уровней, позволяет быстро уменьшить число конкурирующих вариантов, вместо их экспоненциального роста, если используется только информация о прошлых состояниях.

В [28] было показано, что слова, по крайней мере русской речи, записанные в фонетическом коде, обладают свойствами так называемых префиксных кодов, у которых ни одно кодовое слово не служит началом другого. Для 2500 наиболее часто встречающихся слов найдено менее 7% слов-префиксов, которые состоят из одно-двухбуквенных союзов, предлогов и местоимений. Основное свойство префиксных кодов состоит в возможности декодирования слитных сообщений, в которых кодовые слова не разделены паузами или специальными символами. Это имеет принципиальное значение для распознавания слитной речи. Одновременно выяснилось, что вероятность появления фонем в речи определяется не их помехоустойчивостью, а сложностью их образования.

Используя теорему о кодировании и результаты психоакустических экспериментов по восприятию речи в присутствии белого шума с разным отношением сигнал/шум, в [28] была также оценена потенциальная надёжность распознавания слов в случае, когда не используется синтаксическая, семантическая и прагматическая избыточность речи. Оказалось, что при достаточно хороших отношениях сигнал/шум реальная словесная разборчивость и теоретические оценки близки, независимо от того, выполняется ли декодирование по независимым признакам или по сложным комплексам признаков, которыми являются фонемы. При более высоких уровнях шумов теоретически достижима меньшая ошибка распознавания, но человек почему-то не использует все возможности для коррекции ошибок на словесном уровне. Похоже, что при плохих условиях восприятия человек либо использует корректирующую способность более высоких уровней, либо прибегает к переспросу. Это может быть связано с какими-то ограничениями на сложность переработки информации в мозге человека. Аналогичные явления наблюдаются и в технических системах. Например, ограничения на сложность декодирования могут привести к тому, что неоптимальный метод декодирования, не использующий полностью кодовую избыточность, обеспечивает меньшую ошибку, чем метод, потенциально способный использовать всю кодовую избыточность для исправления ошибок, но требующий чрезмерного количества вычислений.

Но с другой стороны, полученные оценки могут свидетельствовать о том, что системы автоматического распознавания речи способны достигнуть гораздо большей устойчивости к аддитивным шумам при достаточных вычислительных ресурсах.

Признаки фонем разделяются на две группы. В одну из них входят признаки, сравнительно легко вычисляемые на акустическом уровне. Это признаки голосового и

шумового источников возбуждения, смычки и назальности. При хороших акустических условиях эти признаки обеспечивают достаточно высокую различимость слов и в совокупности с избыточностью высших уровней гарантируют приемлемую раборчивость речи. Эти же признаки эффективно работают и при быстрой сортировке эталонов больших словарей [29].

Как уже упоминалось выше, имеются экспериментальные свидетельства того, что при высоком уровне помех человек прибегает к вычислению каких-то артикуляторных компонент для улучшения надёжности восприятия речи. Похоже, что для определения места артикуляции действительно необходимо решение обратной задачи относительно формы речевого тракта. Потребность в таком решении тем выше, чем менее доступна информация о синтаксисе, семантике и прагматических ограничениях в задаче понимания речи.

Таким образом, пользуясь детекторами артикуляторных событий для сегментации речевого сигнала и решением обратной задачи для определения места артикуляции, можно ожидать существенного улучшения надёжности распознавания речи.

9. Сжатие речевого сигнала

Системы мобильной связи достигли массового распространения вследствие успехов в сжатии речевого сигнала. Однако, по мере снижения скорости передачи, узнаваемость индивидуальных характеристик голоса снижается и резко падает при скоростях ниже 9 бит/с. Решение обратной задачи относительно команд управления моделью речеобразования позволит снизить предельную скорость в несколько раз.

В наших экспериментах по решению обратной задачи все операции выполнялись с двойной точностью в режиме с плавающей запятой, что соответствует практически непрерывному представлению. Были проведены эксперименты по оценке погрешности артикуляторных параметров при квантовании артикуляторных параметров на 4–6 бит *(табл. 1)*.

Таблица 1

Ошибки аппроксимации (%) траекторий артикуляторов в зависимости от числа уровней квантования

Число бит *N*	Губы	Зубы	Язык — твёрдое нёбо	Язык — мягкое нёбо	Язык — задняя стенка	Нёбная занавеска
4	3,02	3,54	29,32	8,74	14,20	2,76
5	2,09	3,07	16,82	4,37	11,51	1,42
6	1,52	2,79	8,33	3,63	10,41	0,76
7	1,49	2,77	4,43	2,72	6,54	0,59
8	1,46	2,75	2,82	2,88	5,66	0,52
непрерывно	1,44	2,74	2,06	1,72	3,74	0,5

Как видно, квантование на 7–8 бит обеспечивает точность аппроксимации, сопоставимую с «непрерывным» описанием команд. Интервал между сменой команд управления в среднем равнялся примерно 60 мс. Тогда для артикуляторной модели с 16 параметрами и кусочно-линейными управлениями для объективно точной передачи динамики артикуляции потребовалась бы скорость передачи около 2133 бит/сек.

В другой серии экспериментов выполнялась оценка спектральных характеристик и субъективного качества речи при квантовании управлений на 7 бит. Оказалось, что при скорости передачи около 1,8 кбит/с качество синтезированной речи практически не отличается от качества исходного речевого сигнала. Качество синтезированной речи оказалось лучше, чем у стандартного CELP кодера на 9,6 бит/с. Этот результат был получен без использования статистических приёмов, таких как векторное квантование, и без опоры на свойства слуха. Использование этих приёмов должно уменьшить скорость передачи.

Следует отметить, что в этих экспериментах не решалась обратная задача относительно импульса источника голосового возбуждения, и даже при этом индивидуальность голоса воспроизводилась вполне удовлетворительно. Успешные эксперименты по идентификации параметров голосового источника [77, 78] позволяют надеяться на дальнейшее снижение скорости передачи без ухудшения качества синтезированной речи.

Схема артикуляторного вокодера выглядит следующим образом. На передающем конце канала связи решается обратная задача относительно команд управления артикуляцией и параметров голосового источника, эти команды и параметры передаются по каналу связи, а на приёмном конце речевой сигнал синтезируется с помощью модели речеобразования. Для разработки такого вокодера необходимо создать кодовую книгу команд управления и параметров кодового источника. Кроме того, должны быть найдены устойчивые методы анализа формантных частот, что, как известно, до сих пор не реализовано.

10. Верификация диктора

Существуют две задачи распознавания диктора, которые сильно различаются как по постановке, так и по достижимым результатам. Задача идентификации диктора решается с довольно низкой надёжностью порядка 80% в лучшем случае. В этой задаче объём и состав обучающей выборки не обязательно совпадают с условиями распознавания. К тому же диктор не всегда заинтересован в том, чтобы его идентифицировали.

Задача верификации, т.е. подтверждение личности, может быть решена с гораздо большим успехом, поскольку диктор заинтересован в том, чтобы его опознали, и возможно создание обширной базы данных параметров голоса диктора в период обучения. Верификация диктора в настоящее время востребованна во многих областях, таких как санкционирование доступа к компьютерным (включая Интернет) ресурсам или доступа в помещение, сейф, разрешение на запуск двигателя автомобиля, подтверждение права распоряжаться кредитной картой или банковским счётом.

Поскольку в задачах распознавания диктора требуется высокая точность вычисления параметров голоса, то современные методы анализа речи вроде скрытых марковских моделей мало пригодны в силу их неустойчивости к помехам и искажениям речевого сигнала. В этом отношении методы анализа динамических детекторов и решения обратных задач обладают значительным преимуществом.

Один из вариантов верификации заключается в использовании пароля в виде случайной последовательности слов из фиксированного словаря.

Разработка системы голосовой верификации диктора и оценка её эффективности должны производиться в условиях, максимально близких к условиям реальной эксплуатации. Это означает, что помимо достаточно представительного множества дикторов должны использоваться разнообразные типы приёмников звука и аналого-цифровых преобразователей, а запись звука должна производиться в различных помещениях и при различных видах и уровнях посторонних шумов.

Эти требования были выполнены путём формирования специальной базы данных, в которой голоса различных групп дикторов записывались в различных условиях и через различные типы микрофонов и АЦП. В общей сложности, для записи речевых сигналов использовали два типа телефонных трубок и 7 типов микрофонов (направленных, всенаправленных, кардиоидных, с шумоподавлением и без него), размещённых на разных расстояниях от диктора. База данных содержала 429 дикторов (243 мужчины и 186 женщин). Было проведено около 30 миллионов испытаний, так что полученные оценки вероятности ошибки состоятельны. Доверительный интервал точности оценки составлял при этом менее ±0,001%.

Словарь состоял из числительных русского языка от 0 до 9, произносимых по подсказке компьютера *(рис. 10)*.

Ниже показана вероятность (в процентах) суммарной ошибки ложного пропуска и ложного отказа, определённая при работе системы с критерием минимума этой ошибки для дикторов-мужчин. При этом выяснилось, что для паролей длиной в 9 и 10 слов суммарная ошибка примерно поровну делится между ошибками ложного пропуска и ложного отказа.

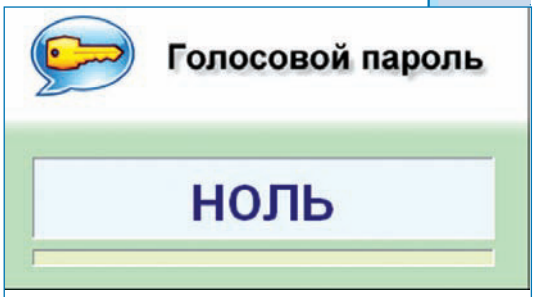

Рис. 10. Подсказка к произнесению пароля

Таблица 2

Ошибка верификации в зависимости от числа слов в пароле для мужских голосов

Число цифр в пароле	1	2	3	4	5	6	7	8	9	10
Ошибка, %	8,49	2,58	0,97	0,43	0,18	0,10	0,054	0,021	0,013	0,006

В *табл. 3* показана доля дикторов-мужчин, имеющих ошибку не более, чем указанную величину.

Таблица 3

Распределение ошибок по мужским голосам

Предельное значение суммарной ошибки (%)	0,001	0,01	0,05	0,25	0,50
Число дикторов с такой ошибкой (%)	73	17	7	3	0

В *табл. 4* показано распределение ошибок по количеству слов в пароле для женщин.

Таблица 4

Ошибка верификации в зависимости от числа слов в пароле для женских голосов

Число цифр в пароле	1	2	3	4	5	6	7	8	9	10
Ошибка, %	10,92	3,70	1,52	0,66	0,35	0,20	0,097	0,060	0,041	0,025

В *табл. 5* показана доля дикторов-женщин, имеющих ошибку не более, чем указанную величину.

Таблица 5

Распределение ошибок по женским голосам

Предельное значение суммарной ошибки (%)	0,001	0,01	0,05	0,25	0,50
Число дикторов с такой ошибкой (%)	72	11	11	3	3

Таким образом, достоверные оценки суммарной ошибки при длине пароля в 10 слов для мужских голосов составляют около 0,006%, а для женских голосов — около 0,025%, но более чем для 70% голосов, как мужчин, так и женщин, гарантируется ошибка менее 0,001%, а для 90% мужчин и 83% женщин гарантируется ошибка менее 0,01%.

Испытание разработанной системы верификации показало, что она устойчива к стационарным шумам с отношением сигнал/шум порядка +10 дБ, а также к посторонним разговорам и музыке.

11. Заключение

Новые измерительные методики, такие как регистрация движений артикуляторов внутри речевого тракта с помощью микролучевого рентгеноскопа и магнитно-резонансная томография (MRI) в трёх измерениях, предоставили исходные данные для дальнейшего развития математических моделей акустики речеобразования, а также позволили поставить на экспериментальную основу исследование возможности решения обратных задач относительно формы речевого тракта, положений артикуляторов и команд управления. Полученные результаты дают основание полагать, что и система управления артикуляцией, и система восприятия речи используют так называемую внутреннюю модель, основанную на способности к решению речевых обратных задач.

Концепция детекторов спектрально-временных неоднородностей была реализована в виде единой математической модели, при разных параметрах которой воспроизводятся эффекты латерального торможения, эффекты нарастания и спада сигнала (on— и off-), эффекты частотных и амплитудных модуляций на слоговом уровне.

Разработанный комплекс математических моделей речеобразования и восприятия, включающий методы решения речевых обратных задач, открывает принципиально новые возможности в решении задач речевой технологии — синтеза речи по тексту, распознавания речи и диктора, сжатия речевого сигнала. В частности, доказана возможность создания артикуляторного вокодера со скоростью передачи ниже 2 Кб/с при практически полном сохранении индивидуальности голоса. Система верификации диктора для фиксированного словаря из десяти числительных русского языка достоверно обеспечивает ошибку менее 0,01%, т.е. на два порядка меньшую, чем у лучших известных систем.

Литература

1. *Баден П., Макаров И.С., Сорокин В.Н.* Алгоритм вычисления площадей поперечных сечений речевого тракта. Акустический ж., 2005. Т. 51. № 1. С. 52–58.

2. *Макаров И.С., Сорокин В.Н.* Резонансы разветвлённого речевого тракта с податливыми стенками, Акустический ж., 2004. Т. 50. № 3. С. 389–396.

3. *Stevens, K.N.* Toward a model for speech recognition, J. Acoust. Soc. Am., 1960. V. 32. P. 47–55.

4. *Galunov, V.I., Chistovich, L.A.* Relationship of motor theory to the general problem of speech recognition. Soviet Physics Acoustics, 1966. V. 11(4). P. 357–365.

5. *Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy M.* Perception of speech code. Physiological review, 1967. V. 74. № 6. P. 431–461.

6. *Liberman, A., Mattingly, I.* The motor theory of speech perception revised. Cognition, 1985. V. 21. P. 1–36.

7. *Fowler, C.A.* An event approach to the study of speech perception from a direct-realist perspective. J. Phonetics, 1986. V.14 (1). P. 3–28.

8. *Ohala, J.J.* Speech perception is hearing sound, not tongues. J. Acoust. Soc. Am., 1996. V. 99. №. 3. P. 1718–1725.

9. *O'Shaughnessy, D.* Critique: Speech perception: Acoustic or articulatory? J. Acoust. Soc. Am., 1996. V. 99. №. 3. P. 1726–1729.

10. *Remez R.E.* Critique: Auditory form and gestural topology in the perception of speech. J. Acoust. Soc. Am., 1996. V. 99. №. 3. P. 1695–1698.

11. *Lindblom, B.* Role of articulation in speech perception: Clues from production. J. Acoust. Soc. Am., 1996. V. 99. №. 3. P. 1683–1692.

12. *Stevens, K.N.* Critique: Articulatory-acoustic relations and their role in speech perception. J. Acoust. Soc. Am., 1996. V. 99. №. 3. P. 1693–1694.

13. *Lindblom, B.E.F., Lubker, J., and Gay, T.* Formant frequencies of some fixed mandible vowels and a model of speech motor programming by predictive simulations. J. Phonetics, 1979. V. 7. P. 147–161.

14. *Fowler, C.A., Turvey, M.T.* Immediate compensation in bite-block speech. Phonetica, 1980. V. 37. P. 306–326.

15. *Gay, T.J., Lindblom, B., and Lubker, J.* Production of bite-block vowels: Acoustic equivalence by selective compensation. J. Acous. Soc. Am., 1981. V. 69. P. 802–810.

16. *Flege, I.E., Fletcher, S.G., and Homiedan, A.* Compensating for a bite block in /s/ and /t/ production: Palatographic, acoustic, and perceptual data. J. Acoust. Soc. Am., 1988. V. 83. № 1. P. 212–228.

17. *Savariaux, C., Perrier. P., and Orliaquet, J.P.* Compensation strategies for the perturbation of the rounded vowel [*u*] using a lip tube: A study of the control space in the speech production. *J.* Acoust. Soc. Am., 1995. V. 98. №. 5. P. 2428–2442.

18. *Hamlet, S.L., Stone, M.* Compensatory alveolar consonant production induced by wearing a dental prosthesis. J. Phonetics, 1978. V. 6. P. 227–248.

19. *MacFarland, D.H., Baum, S.R.* Incomplete compensation to articulatory perturbation. J. Acoust. Soc. Am., 1995. V. 97. № 3. P. 1865–1873.
20. *Munhall, K.G., Lofqvist, A., and Kelso J.A.S.* Lip-larynx coordination in speech: Effects of mechanical perturbations to the lower lip. J. Acoust. Soc. Am., 1994. V. 95. № 6. P. 3605–3616.
21. *Sorokin, V., Olshansky, V., and Kozhanov L.* Internal model in articulatory control: Evidence from speaking without larynx. Speech Communication, 1998. V. 19. № 3. P. 249–268.
22. *Burnett, T.A., Freedland, M.B., Larson, C.R., and Hain, T.C.* Voice F_0 responses to manipulations in pitch feedback, J. Acoust. Soc. Am., 1998. V. 103. P. 3153–3161.
23. *Jones, J.L., Munhall, K.G.* Perceptial calibration of F_0 production: Evidence from feedback manipulation. J. Acoust. Soc. Am., 2000. V. 108. P. 1246–1251.
24. *Max, L., Wallace, M.E., Vincent, I.* Sensorimotor adaptation to auditory perturbations during speech: Acoustic and kinematic experiments. Proc. 15th ICPhS, Barcelona, 2003. V. 1. P. 1053–1056.
25. *Villacorta, V., Perkell, J.S., and Guenter, F.H.* Sensorimotor adaptation to acoustic perturbations in vowel formants. J. Acoust. Soc. Am., 2004. V. 115. P. 2618.
26. *Purcell, D.W., Munhall, K.G.* Compensation following real-time manipulation of formants in isolated vowels. J. Acoust. Soc. Am., 2005. V. 119(4). P. 2288–2297.
27. *Niyogy. P., Sondhi, M.M.* Detecting stop consonants in continuous speech. J. Acoust. Soc. Am., 2002. V. 111. P. 1063–1076.
28. *Сорокин В.Н.* Теория речеобразования. М.: Радио и связь, 1985.
29. *Sorokin V.N.* Some coding properties of speech. Speech Communication, 2003.V. 40. № 3. P. 409–423.
30. *Callan, D.E., Callan, A.M., Kroos, Ch., and Vatikiotis-Bateson, E.* Neural processes underlying perception of audio-visual speech production. Proc. 5th Seminar on Speech Production, Kloster Seeon. 2000. P. 273–276.
31. *Sekiyama, K., Sugita, Y.* Auditory-visual speech perception examined by brain imaging and reaction time. Proc. 7th Int. Conf. On Spoken Language, 2002.
32. *Sams, M., Aulanko, R., Hamalainen, H., Lounasmaa,O., Lu S., and Simola, J.* Visual information from lip movements modifies activity in the human auditory cortex. Neuroscience Letters, 1991. V. 127. P. 141–145.
33. *Calvert, G., Brammer, M., Bullmore, E., Campbell, R., Williams, S., McGuire. P., Voodruff. P., Iversen, S., and David,A.* Activation of auditory cortex during silent lip-reading. Science, 1997. V. 276. P. 593–596.
34. *Folkins, J.W., Abbs, J.H.* Lip and jaw motor control during speech: Responses to resistive loading of the jaw. J. Hearing and Speech Res., 1975. V. 18. P. 207–220.
35. *Abbs, J.H., Gracco, V.L.* Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech, J. Neurophysiology, 1984. V. 51. P. 705–723.
36. *Kelso, J.A.S., Tuller, B., Vatiokis-Bateson, E., and Fowler, C.* Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. J. of Experimental Psychology: Human Perception and Performance, 1984. V. 10. P. 812–832.
37. *Shaiman, S.* Kinematic and electromyographic responses to perturbation of the jaw. J. Acoust. Soc. Am., 1989. V. 86. P. 78–87.
38. *Kollia, H.B., Gracco, V.L., and Harris, K.S.* Functional organization of velar movements following jaw perturbation. J. Acoust. Soc. Am., 1992. V. 91, p. 2474.
39. *Honda, M., Kaburagi, T.* Speech compensation to dynamical structural perturbation of the palate shape. In: Proc. 5th Seminar on Speech Production, Kloster Seeon, Bavaria, 2000. P. 21–24.
40. *Folkins, J.W., Zimmerman, G.N.* Lip and jaw interactions during speech: responses to

perturbation of lower-lip movement prior to bilabial closure. J. Acoust. Soc. Am., 1982. V. 71. № 4. P. 1225–1233.
41. *Abbs, J.H., Cole K.J.* Consideration of bulbar and suprabulbar afferent influences upon speech motor control coordination and programming // *Speech Motor Control*, Pergamon Press, London, 1982. P. 159–186.
42. *Lee B.S.* Effects of delayed speech feedback. J. Acoust. Soc. Am., 1950. № 22. P. 824.
43. *Poeck, K., Orgass, B.* The concept of body scheme, a critical review and experimental results. Cortex, 1971. V. 5. P. 254–277.
44. *Kelso, S.J., Stelmach, G.E.* Central and peripheral mechanisms in motor control // *Motor Control. Issues and Trends,* Stelmach, G.E., (Ed.), Academic Press, N-Y., 1976. P. 3–40.
45. *Weinstein, S., Sersen, E.A.* Phantoms in cases of congenital absence of limbs. Neurology, 1961. V. l. № 10–11. P. 905–911.
46. *Vetter, R.J., Weinstein, S.* The history of the phantom in congenital absent limbs. Neuropsychology, 1967. V. 5. P. 335–338.
47. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. М. Наука, 1974.
48. *Леонов А.С., Сорокин В.Н.* Обратная задача для управления артикуляцией, Доклады Академии Наук, 2000. Т. 374. № 6. С. 749–753.
49. *Леонов А.С., Макаров И.С., Сорокин В.Н., Цыплихин А.И.* Артикуляторный ресинтез гласных, Информационные процессы, 2003. Т. 3. № 2. С. 73–92.
50. *Atal, B.S., Chang, J.J., Mathews, M.V., Tukey, J.W.* Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique, J. Acoust. Soc. Am., 1978. V. 63. P. 1535–1555.
51. *Sorokin, V.N.* Determination of vocal tract shape for vowels. Speech Communication, 1992. V. 11. № 1. P. 71–85.
52. *Schmidt R.A.* More on motor programs // *Human Motor Behavior,* Kelso, S., (Ed.), Lawrence Erlbaum Associates, Hillsdale, New Jersey. 1982. P. 189–218.
53. *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. V. 10. P. 110–126.
54. *Larar, J.N., Schroeter, J., and Sondhi, M.M.* Vector-quantization of the articulatory space, IEEE Trans. Acoust., Speech and Signal Processing, ASSP-36, 1988. № 12. P. 1812–1818.
55. *Schroeter, J., Meyer, R., and Parthasarathy, S.* Evaluation of improved articulatory codebooks and codebook access distance measures. Proc. Internat. Conf. Acoust. Speech Signal Processing, 1990. P. 393–396.
56. *Schroeter, J., Sondhi, M.M.* Dynamic programming search of articulatory codebooks. IEEE Proc. Int. Conf. Acoust. Speech Signal Processing, 1989. P. 588–591.
57. *Sorokin V.N., Trushkin A.V.* Articulatory-to-acoustic mapping for inverse problem. Speech Communication, 1996. V. 19. № 4. P. 105–118.
58. *Леонов А.С., Макаров И.С., Сорокин В.Н., Цыплихин А.И.* Кодовая книга для речевых обратных задач. Информационные процессы, 2005. Т. 5. № 2. С. 101–119.
59. *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. V. 30, № 1. P. 55–74.
60. *Sumby, W.Y., Pollack, I.* Visual contribution to speech intelligibility in noise, J. Acoust. Soc. Am., 1954. V. 26. P. 212–215.
61. *McGurk, H., MacDonald, J.* Hearing lips and seeing voices, Nature, 1976. V. 264. P. 746–748.
62. *Westbury J.* X-ray Microbeam Speech Production Database. User's Handbook, Version 1, 1994.
63. *Sorokin V.N.* Inverse problem for fricatives. Speech Communication, 1994. V. 14. № 2. P. 249–262.
64. *Леонов А.С., Макаров И.С., Сорокин, В.Н. Цыплихин А.И.* Артикуляторный ресинтез фрикативных. Информационные процессы, 2004. Т. 4. № 2. С. 141–159.
65. *Leonov A.S, Sorokin V.N.* Optimality criteria in inverse problems for tongue-jaw interaction, 2003, Proc. EuroSpeech, 2003. P. 2353–2356.
66. *Leonov A.S., Sorokin V.N.* Control in the Internal Model: Score Reorganization and Compensation, Pattern Recognition and Image Analysis, 2004. V. 14. № 3. P. 407–420.
67. *Kent, R.D., Carney. P.J., and Severeid L.R.* Velar movement and timing: evaluation of a model for binary control. J. of Speech and Hearing Research, 1974. V. 17. P. 470–488.
68. *Gay, T., Ushijima, T., Hirose, H., and Cooper, F.* Effect of speaking rate on labial and consonant-vowel

articulation. J. Phonetics, 1974. V. 2. P. 47–63.
69. *Dau T., Puschel D., Kohlrausch A.* A quantitative model of the «effective» signal processing in the auditory system: I. Model structure // Journ. Acoust. Soc. Am., 1996. V. 99. № 6. P. 3615–3622.
70. *Lippmann R.P.* Speech recognition by machines and humans // Speech Communication, 1997. V. 22. № 1. P. 1–16.
71. *Kingsbury B.E.D., Morgan N., Greenberg S.* Robust speech recognition using the modulation spectrogram // Speech Communication. 1998. V. 25. № 1–3. P. 117–132.
72. *Hermansky H.* Perceptual linear predictive (PLP) analysis of speech // Journ. Acoust. Soc. Am. 1990. V. 87. № 4. P. 1738–1752.
73. *Tchorz J., Kollmeier D.* A model of auditory perception as front end for automatic speech recognition, JASA, 1999. 106(4), Pt.1. P. 2040–2050.
74. *Сорокин В.Н., Чепелев Д.Н.* Первичный анализ речевых сигналов. Акустический ж., 2005. Т. 51. № 4. С. 536–542.
75. *Сорокин В.Н.* Синтез речи. М.: Наука, 1992.
76. *Martin, A., Fiscus, J., Przybocki, M., Fisher, B.* The evaluation: word error rates and confidence analysis. Proc. 9th Hub-5 Conversational Speech Recognition Workshop, Linthicum Heights, MD. 1998.
77. *Сорокин В.Н., Макаров И.С.* Обратная задача для голосового источника. Информационные процессы. 2006. Т. 6. № 4. С. 375–395.
78. *Сорокин В.Н., Макаров И.С.* Распознавание пола диктора по голосовому источнику.

Сорокин Виктор Николаевич,

Доктор физико-математических наук, ведущий научный сотрудник Института проблем передачи информации РАН. Занимается фундаментальными исследованиями процессов речеобразования и восприятия речи и приложениями к речевым технологиям с 1964 г. Опубликовал более 130 работ, в том числе монографии «Теория речеобразования», 1985, «Радио и Связь» и «Синтез речи», (М.: Наука, 1992).