

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

## Обратная задача для голосового источника

В.Н.Сорокин, И.С.Макаров

*Институт проблем передачи информации, Российская академия наук, Москва, Россия*  
Поступила в редколлегию 19.12.2006

### Аннотация

Исследовалась обратная задача относительно формы голосового источника. Входными параметрами для этой задачи служили либо сигнал-остаток, получаемый после обратной фильтрации речевого сигнала, либо текущий интеграл от этого остатка. Сам сигнал-остаток интерпретируется как производная от объемной скорости воздушного потока через голосовую щель, а его интеграл – как объемная скорость потока. Рассматривалось несколько способов решения обратной задачи. В одном из них обратная задача решалась вариационным методом с регуляризацией по Тихонову. В этом методе параметры модели колебаний голосовых складок варьировались с целью минимизации целевого функционала, включающего в себя невязку между вычисленной по модели производной от объемной скорости и сигналом, полученным после обратной фильтрации.

В другом методе сначала выполнялась оценка объемной скорости по сигналу обратной фильтрации, а затем эта объемная скорость использовалась для решения инвертированного уравнения динамики потока в голосовом источнике. Полученная оценка изменения площади голосовой щели на каждом периоде голосового источника аппроксимировалась по методу среднеквадратического минимума функцией, порождаемой моделью голосового источника. Этот метод оказался наиболее точным и устойчивым. Ошибки аппроксимации известной площади голосовой щели в этом методе находились в пределах до 0.1 % (синтетические гласные), 10 – 12 % (натуральные гласные с измеренной площадью голосовой щели).

Установлено, что правдоподобные оценки площади голосовой щели достигаются на близко расположенных микрофонах.

### 1. Введение

Параметры колебаний голосовых складок, определяющие форму импульсов голосового возбуждения в речевом тракте, могут быть полезны как в задачах речевых технологий, таких как идентификация и верификация диктора по голосу или автоматическое распознавание речи, так и при оценке эмоционального или физиологического состояния диктора, в том числе и для диагностики заболеваний гортани.

Исходными данными для определения параметров голосового источника служат результаты так называемой обратной фильтрации речевого сигнала. Эта технология основана на предположении линейности системы "голосовой источник – речевой тракт", т.е. на отсутствии влияния акустических характеристик речевого тракта на колебания голосовых складок. В таком случае амплитудно-частотная характеристика речевого сигнала определяется произведением передаточных функций голосового источника и речевого тракта. Если бы каким-либо образом удалось оценить параметры передаточной функции речевого тракта, то можно было бы вычислить и параметры голосового источника.

Голосовой источник возбуждения акустических колебаний в речевом тракте создается силой, пропорциональной производной от объемной скорости воздушного потока, протекающего через голосовую щель. Автоматическая оценка параметров голосового источника предполагает последовательное решение следующих задач:

1. вычисление первой производной  $w'(t)$  от объемной скорости по речевому сигналу (задача обратной фильтрации),
2. определение объемной скорости  $w(t)$  путем интегрирования ее производной,
3. аппроксимация функции  $w(t)$  или  $w'(t)$  некоторой моделью голосового источника,
4. вычисление площади голосовой щели  $S(t)$ .

В мировой литературе решению этих задач посвящено значительное количество работ.

Одной из первых работ, посвященных решению задачи **обратной фильтрации**, стала работа [1]. Метод решения заключался в оценке полюсов речевого тракта по измеренному акустическому сигналу и дальнейшем пропускании этого сигнала через фильтр, передаточная функция которого имела только нули, совпадающие с полюсами речевого тракта. При этом предполагалось, что передаточная функция речевого тракта содержит только полюсы, а спектр речевого сигнала есть произведение передаточной функции тракта на спектр голосового источника.

В работе [2] было предложено аппроксимировать передаточную функцию речевого тракта  $A(z)$  (в  $z$ -области) как

$$A(z) = \frac{1}{1 + \sum_{k=1}^p a_k z^{-k}}, \quad (1)$$

причем коэффициенты фильтра  $a_i$  определялись путем анализа речевого сигнала моделью линейного предсказания порядка  $p$ . После определения коэффициентов модели речевой сигнал пропусклся через фильтр с передаточной функцией, равной  $1/A(z)$ . Отклик этого фильтра и служил оценкой  $w'$  через голосовую щель.

Работа [2] оказалась этапной – метод обратной фильтрации с помощью модели линейного предсказания является в настоящее время наиболее популярным. В ряде работ [3 - 6] для повышения точности обратной фильтрации предложено использовать данные электроглоттограмм. В работах [6 - 11] результаты обратной фильтрации уточнялись с помощью различных многоступенчатых алгоритмов. Работы [12 - 14] посвящены совместному определению коэффициентов линейного предсказания и параметров модели голосового источника. Наконец, некоторые исследователи осуществляли обратную фильтрацию с помощью методов, отличных от метода линейного предсказания (например, с помощью метода дискретной полюсной модели (ДПМ - Discrete All - Pole, DAP) – [15,16], метода кепстрального разложения [17], или метода нечеткой кластеризации [18]).

Все математические модели, используемые для аппроксимации результатов обратной фильтрации, можно разделить на три класса – модели механических колебаний голосовых складок, модели площади голосовой щели и параметрические модели объемной скорости или ее первой производной. Ввиду простоты программной реализации именно последний класс моделей получил наибольшую популярность в работах по обратной фильтрации.

По поводу использования механических моделей голосовых складок или параметрических моделей площади голосовой щели можно сделать следующее общее замечание: во всех известных работах параметры этих моделей подбираются путем минимизации невязки (в некоторой метрике) между результатами обратной фильтрации и голосовым источником, вычисленным по модели. В теории некорректных задач [19] показано, что в общем случае такой подход приводит к решениям, неустойчивым относительно малых возмущений данных задачи. Вместо этого рекомендуется использовать так называемые регуляризирующие алгоритмы, обеспечивающие существование, единственность и устойчивость решения.

Основная цель данной работы заключается в создании полностью автоматического устойчивого алгоритма оценки параметров голосового источника. Достижение этой цели требует решения ряда задач:

- сопоставление различных алгоритмов обратной фильтрации и выбор наилучшего (относительно выбранной цели) алгоритма;
- создание адекватной параметрической модели площади голосовой щели;
- построение алгоритма регуляризации данных обратной фильтрации с помощью модели площади голосовой щели;

Все эти задачи решаются в последующих разделах этой работы. Сопоставление различных методов обратной фильтрации и анализ точности построенного алгоритма проводится на материале гласных звуков, синтезированных для различных конфигураций голосовой щели и различной частоты основного тона, а также на материале синхронных измерений акустического сигнала и функции площади голосовой щели с помощью высокоскоростной киносъемки. Эти синхронные измерения были любезно предоставлены нам проф. Чилдерсом (Donald G. Childers, University of Florida, Gainesville).

## 2. Параметрическая модель голосового источника

Уравнение для аэродинамического потока через голосовую щель было описано в [20]. Это нелинейное дифференциальное уравнение первого порядка типа уравнения Риккати:

$$\rho_0 h(vS)' + k_{mp} h v S + \frac{c_x \rho_0}{2} S v^2 = \Delta p S. \quad (2)$$

Здесь  $v$  – скорость потока,  $S$  – площадь голосовой щели,  $\rho_0$  – плотность воздуха,  $h$  – глубина голосовой щели вдоль оси потока,  $\Delta p$  – перепад давления над голосовой щелью,  $c_x$  – коэффициент динамического сопротивления, который зависит от формы голосовой щели и числа Рейнольдса. Штрих означает производную по времени. Коэффициент вязкого трения  $k_{mp}$  рассчитывается как для капиллярной трубки прямоугольного сечения с наименьшим размером  $b$ :

$$k_{mp} = 12\mu b^2 S^2,$$

$\mu$  – коэффициент вязкости воздуха,  $\mu = 1.86 \times 10^{-4}$  г/см·с.

В [21] получена рекурсивная форма этого нелинейного уравнения, представляющая собой алгебраическое уравнение второго порядка относительно объемной скорости воздушного потока  $w = vS$ :

$$\alpha\beta c_x \rho_0 w^2(t + \Delta t) - 2[w(t)(1 - \alpha) - w(t + \Delta t)]S(t + \Delta t) - 2\alpha\beta\Delta p(t + \Delta t)S^2(t + \Delta t) = 0, \quad (3)$$

где  $\Delta t$  – период дискретизации по времени и

$$\alpha = 1 - e^{-\Delta t/T}; \quad \beta = \frac{T}{\rho_0 h}; \quad T = \frac{\rho_0 h}{k_{mp}}.$$

Уравнение (3) можно решать как относительно объемной скорости  $w$ , так и относительно площади голосовой щели  $S$ . Задавая геометрическими параметрами голосовой щели, перепадом давления и формой изменения площади голосовой щели во времени, получаем решение для объемной скорости потока через голосовую щель:

$$w(t + \Delta t) = \frac{\sqrt{1 + 4a_1 a_2} - 1}{2a_2}, \quad (4)$$

где

$$a_1 = w(t) + \alpha[\beta\Delta p(t + \Delta t) - w(t)],$$

$$a_2 = \frac{\alpha\beta c_x \rho_0}{2S(t + \Delta t)}.$$

Решение (4) оказалось достаточно точным для тех параметров, которые реально используются в исходном дифференциальном уравнении (2) для голосовой щели. Производная от объемной скорости  $w$  создает голосовой источник, который обеспечивает весьма высокую натуральность синтезированной речи при условии правильного выбора класса функций, описывающего динамику изменения площади голосовой щели  $S(t)$ .

Производная  $w'$  может быть вычислена непосредственно из уравнения (2) после того, как из (4) найдена объемная скорость  $w$ :

$$w'(t) = \frac{\Delta p S(t)}{\rho_0 h} - \frac{c_x}{2hS(t)} w^2(t) - \frac{12\mu b^2 S^2(t)}{\rho_0} \quad (5)$$

Вместе с тем, использование левой конечной разности для вычисления производной

$$w'(t) = [w(t) - w(t + \Delta t)] / \Delta t \quad (6)$$

показало, что, по сравнению с (5), погрешность достаточно мала, тогда как процесс вычисления значительно проще.

Автоколебания голосовых складок являются результатом взаимодействия трехмерных упругих деформаций тканей складок и эффекта Бернулли, создаваемого аэродинамическим потоком через голосовую щель. Эти процессы описываются сложной системой нелинейных дифференциальных уравнений,

решение которых занимает слишком большое время. Поэтому в речевых исследованиях используют упрощенные модели, описывающие импульс объемной скорости, его производную по времени, либо изменения площади голосовой щели во времени. Модели производной от объемной скорости удобны тем, что они могут непосредственно использоваться при аппроксимации сигнала, полученного в результате обратной фильтрации [6, 8, 14, 15, 22 - 27]. Наиболее популярной является так называемая LF-модель, построенная в [28]. Однако ошибки аппроксимации и дисперсия оценок параметров голосового источника оказываются слишком большими для некоторых дикторов и некоторых типов голосов [6, 23, 25]. В ряде работ объемная скорость через голосовую щель или ее первая производная аппроксимировалась либо степенными [12, 29, 30], либо тригонометрическими полиномами [31, 32], либо решением некоторой автономной системы обыкновенных дифференциальных уравнений [33, 34]. Основным недостатком таких моделей заключается в невозможности физической и физиологической интерпретации многих ее параметров.

В работах [35 – 38] результаты обратной фильтрации аппроксимировались голосовым источником, порожденным параметрическими моделями площади голосовой щели. Такие модели выгодно отличаются как от параметрических моделей голосового источника (физиологической интерпретируемостью всех параметров модели и физической адекватностью соответствующей функции объемной скорости), так и от механических моделей голосовых складок (простотой реализации и вычислительной устойчивостью).

В предварительных экспериментах мы использовали математическую модель площади голосовой щели, описанную в [21, 39]. Эта модель определяется тремя параметрами:  $T_0$  – текущий период основного тона,  $t_1$  – отношение фазы открытия голосовой щели к  $T_0$ ,  $t_2$  – отношение интервала открытой голосовой щели к периоду  $T_0$ . Наши эксперименты показали, что эта модель должна быть дополнена еще двумя параметрами, определяющими скорость раскрытия и закрытия голосовой щели. На необходимость учета подобных параметров указывалось и в [40].

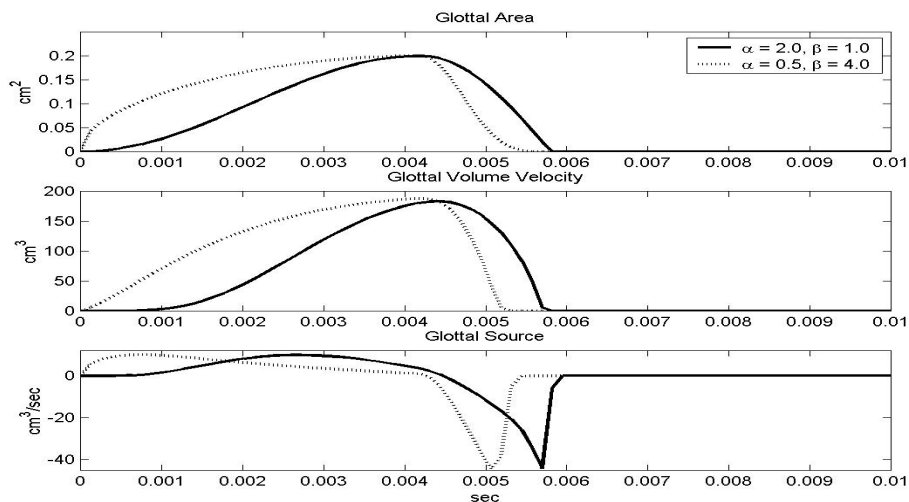


Рис. 1. Площадь голосовой щели (верхний рисунок), соответствующая объемная скорость (средний рисунок) и первая производная от объемной скорости (нижний рисунок).

Модифицированная модель площади голосовой щели  $S(t)$  как функция времени на интервале одного периода основного тона длительностью  $T_0$  описывается как:

$$S(t) = \begin{cases} S_{\max} \left[ \sin \left( \frac{\pi t}{2t_1 T_0} \right) \right]^\alpha, & 0 \leq t \leq t_1 T_0 \\ S_{\max} \left[ \cos \left( \frac{\pi(t - t_1 T_0)}{2(t_2 - t_1) T_0} \right) \right]^\beta, & t_1 T_0 < t \leq t_2 T_0 \\ 0, & t_2 T_0 < t \leq T_0 \end{cases} \quad (7)$$

Здесь  $S_{max}$  – максимальная площадь открытия голосовой щели,  $\alpha, \beta$  – коэффициенты, определяющие скорость раскрытия и закрытия голосовой щели, соответственно. Эта модель оказалась вполне работоспособной. На рис. 1 показаны формы голосовой щели для разных значений коэффициентов  $\alpha, \beta$ , а также соответствующие функции объемной скорости и голосового источника на одном периоде основного тона.

В отличие от решения уравнения (2) относительно объемной скорости, решение относительно площади голосовой щели не рекурсивно:

$$S(t) = \frac{b \pm \sqrt{1 + 4ac}}{2a}, \quad S(t) \geq 0, \quad (8)$$

где

$$\begin{aligned} a &= 2\alpha\beta\Delta p(t + \Delta t), \\ b &= 2[w(t + \Delta t) - (1 - \alpha)w(t)], \\ c &= \alpha\beta c_x \rho_0 w^2(t + \Delta t). \end{aligned}$$

В действительности, уравнение (8) не точно описывает решение (3) относительно площади голосовой щели, поскольку коэффициент вязкого трения в (3) сам зависит от  $S(t)$ . Однако тестирование на синтезированных сигналах показало, что при удачном выборе фиксированного значения  $S$  при вычислении этого коэффициента решение (8) очень близко к истинной функции  $S(t)$ . Также выяснилось, что достаточно точные решения получаются и при фиксации параметров  $b, h$  и  $\Delta p$ , если задать средние значения этих величин по ранее определенным диапазонам.

В последующих разделах приводятся результаты исследования нескольких способов оценки параметров голосового источника. Первый способ состоит в определении формы импульса голосового источника  $w'_{эксп}$  путем обратной фильтрации речевого сигнала и сравнении его с производной от объемной скорости модели источника, описанной уравнениями (4), (6) и (7). Второй способ состоит в интегрировании экспериментально определенной формы импульса источника  $w'_{эксп}$ , вычислении площади голосовой щели  $S_{эксп}$  по (8) и подгонке к ней методом среднеквадратической минимизации модели (7).

### 3. Алгоритмы обратной фильтрации

Поскольку в каждом методе исходными данными служат оценки производной объемной скорости голосового источника, полученные методом обратной фильтрации, ниже рассматриваются известные по литературе алгоритмы.

В соответствии с предположением о независимости источника возбуждения и речевого тракта, амплитудно-частотная характеристика гласноподобных неназализованных сегментов описывается следующим уравнением [41]:

$$P(z) = R(z)A(z)U(z)$$

Здесь  $P(z)$  – спектр (в  $z$ -области) речевого сигнала,  $A(z)$  – передаточная функция речевого тракта, определяемая как отношение спектра объемной скорости у губ к спектру объемной скорости у голосовой щели;  $U(z)$  – спектр функции  $w$ ,  $R(z)$  – акустический импеданс излучения звука через губы. Принимается, что передаточная функция  $A(z)$  имеет только полюсы и, следовательно, описывается соотношением (1). Кроме того, обычно предполагается, что импеданс излучения через губы с высокой точностью моделируется как

$$R(z) = 1 - z^{-1},$$

так что  $z$ -преобразованием функции  $w'$  оказывается произведение  $R(z)U(z)$ . Отсюда следует, что если полюсы передаточной функции  $A(z)$  известны, то оценкой функции  $w'$  будет обратное  $z$ -преобразование от выражения  $P(z)/A(z)$ .

Таким образом, точность обратной фильтрации напрямую зависит от точности определения полюсов передаточной функции речевого тракта. Практика показывает, что на точность обратной фильтрации также влияет расположение частоты первой форманты относительно частоты основного тона – чем эти частоты ближе, тем ниже точность оценки голосового источника с помощью обратной фильтрации. Наи-

худший случай для обратной фильтрации – это гласный с низкой частотой первой форманты, произнесенный с высокой частотой основного тона (например, гласные /i/ или /u/, произнесенные высокими женскими или детскими голосами). Этот факт хорошо известен в литературе по обратной фильтрации.

Перед обратной фильтрацией над речевыми сигналами производились следующие операции. Сначала сигналы пропускались через линейно-фазовый нерекурсивный ФНЧ с частотой среза 3.8 кГц. Затем речевые сигналы снова фильтровались с помощью нуль-фазового нерекурсивного ФВЧ с частотой среза 20 Гц. Использование такого ФВЧ необходимо для подавления постоянной составляющей в речевом сигнале, а также для подчеркивания различия между фазами открытой и закрытой голосовой щели в голосовом источнике.

При оценке голосового источника с помощью автокорреляционного или ковариационного алгоритмов, асинхронных с ОТ, речевой сигнал сначала предсказался с помощью нерекурсивного ФВЧ первого порядка:

$$Q(z) = 1 - 0.9z^{-1}$$

Длительность временного окна анализа для обоих алгоритмов была равна 25 мс (200 отсчетов при частоте дискретизации 8 кГц). При оценке коэффициентов ЛП по предсказанному сигналу в автокорреляционном методе использовалось окно Хэмминга, а в ковариационном методе – прямоугольное окно. При пропуске непредсказанного сигнала через обратный фильтр в обоих алгоритмах использовалось прямоугольное окно, длительность которого также была равна 25 мс. Порядок модели линейного предсказания для обоих алгоритмов был равен 10.

Параметры всех остальных алгоритмов заимствовались из статей, описывающих эти алгоритмы.

Для проведения полномасштабного сопоставления были программно реализованы 11 алгоритмов обратной фильтрации (Табл. 1). Алгоритмы выбирались с тем условием, чтобы различия в каждой паре алгоритмов были весьма существенными (а не сводились бы к простой модификации одного из членов пары).

Сопоставление различных алгоритмов обратной фильтрации осуществлялось на материале синтетических и реальных гласных, описанных в следующем разделе.

Сначала для этих сигналов определялась  $w'_{эксн}$  с помощью каждого из алгоритмов из Табл. 1. Для сравнительного анализа выбиралось несколько периодов первой производной от объемной скорости. Эти периоды вместе с наложенными на них соответствующими периодами «истинного» голосового источника отображались на мониторе для визуальной оценки качества решения задачи обратной фильтрации. Для количественной оценки использовалось среднее значение (по всем периодам) среднеквадратических ошибок между «истинным» голосовым источником и голосовым источником из обратной фильтрации, вычисленных для каждого периода.

В процессе сопоставления разных алгоритмов на материале синтетических гласных были сделаны следующие наблюдения.

Автокорреляционный и ковариационный алгоритмы, асинхронные с ОТ, оказались весьма успешными при решении задачи обратной фильтрации. Для низкого или среднего ОТ (80 и 100 Гц) эти алгоритмы дали очень малые ошибки оценки голосового источника (2 – 5 %). При повышении основного тона и понижении частоты первой форманты ошибки росли (до 20 %). Зависимости между качеством обоих алгоритмов и конфигурацией голосовой щели обнаружено не было. Также не было выявлено преимущества одного из этих алгоритмов перед другим.

Алгоритм, основанный на CELP, дал неприемлемо большие ошибки в оценке голосового источника (до 40 %) для большинства синтетических гласных. При этом оказалось, что разница между исходным и вычисленным голосовым источником тем больше, чем выше частота ОТ (вне зависимости от частоты первой форманты) и чем короче фаза закрытой голосовой щели. Таким образом, этот алгоритм оказался неработоспособным в большинстве случаев.

Из трех алгоритмов, реализующих двухступенчатую схему ЛП, наилучшим был признан первый. Будучи вычислительно эффективным, он дал весьма малые ошибки (2 – 4 %) практически для всех синтетических гласных. Вместе с тем оказалось, что точность двухступенчатого ЛП – 1 несколько падает с ростом частоты основного тона и уменьшением фазы закрытой голосовой щели (соответствующие ошибки возрастают до 12 %).

Табл. 1. Исследованные алгоритмы обратной фильтрации (ОТ – основной тон, ЛП – линейное предсказание, ДПМ – дискретная полюсная модель).

Алгоритм	Краткое описание	Источник
Автокорреляционный (асинхронный с ОТ)	По предсказанному сигналу определяются коэффициенты ЛП автокорреляционным методом. По ним строится обратный фильтр, через который пропускается непредсказанный сигнал. Выходом является оценка $w'$ .	[41]
Ковариационный (асинхронный с ОТ)	По предсказанному сигналу определяются коэффициенты ЛП ковариационным методом. По ним строится обратный фильтр, через который пропускается непредсказанный сигнал. Выходом является оценка $w'$ .	[29, 41]
Алгоритм, основанный на CELP	Алгоритм заимствует все операции анализа речевого сигнала и интерполяции коэффициентов ЛП из стандарта FS 1016 CELP.	[42]
Двухступенчатое ЛП – 1	На первом шаге функция $w'$ оценивается с помощью автокорреляционного алгоритма, асинхронного с ОТ. На втором шаге коэффициенты ЛП определяются для каждого периода ОТ.	[8]
Двухступенчатое ЛП – 2	Отличие от алгоритма ЛП-1 состоит в том, что на втором шаге окно длительностью в четверть периода ОТ последовательно сдвигается на один отсчет, и выбираются коэффициенты ЛП для сигнала–остатка с наименьшей энергией.	[6]
Двухступенчатое ЛП – 3	Отличие от алгоритма ЛП-1 состоит в том, что на втором шаге длительность окна последовательно увеличивается на один отсчет (внутри текущего периода). В каждом окне определяются коэффициенты ЛП и соответствующая частота первой форманты. Длительность окна выбирается так, что увеличение ее на один отсчет приводит к скачкообразному изменению оценки частоты первой форманты.	[10]
Адаптивный выбор предсказывающего фильтра	Сначала $w'$ оценивается асинхронным автокорреляционным методом. На втором шаге $w'$ аппроксимируется полюсно-нулевым рекурсивным фильтром. После этого к исходному сигналу опять применяется автокорреляционный метод, с той лишь разницей, что теперь сигнал предсказывается фильтром, обратным к фильтру, аппроксимирующему голосовой источник.	[11]
Нечеткая кластеризация	Алгоритм основан на методе нечеткой кластеризации точек в многомерном пространстве на попарно ортогональные гиперплоскости.	[18]
ДПМ-алгоритм, асинхронный с ОТ	Алгоритм аналогичен асинхронному автокорреляционному алгоритму, но вместо автокорреляционного метода коэффициенты обратного фильтра определяются с помощью ДПМ-алгоритма.	[15]
Многоступенчатый ДПМ-алгоритм	Алгоритм реализует итерационную схему последовательного определения передаточной функции речевого тракта и голосового источника, при этом параметры источника и передаточной функции определяются с помощью ДПМ-алгоритма.	[16]
Кепстральное разложение	На каждом периоде ОТ по речевому сигналу вычисляется его комплексный кепстр. Затем каузальная часть этого кепстра зануляется. Обратное преобразование полученного комплексного кепстра во временную область и служит оценкой функции $w$ на текущем периоде ОТ.	[17]

Хорошо известно, что на интервале закрытых голосовых складок речевой сигнал представляет собой линейную комбинацию затухающих гармонических колебаний с частотами и ширинами, равными частотам и ширинам формант [41]. Поэтому ковариационный анализ, проведенный на таком интервале,

мог бы обеспечить высокоточные оценки формантных параметров, а голосовой источник, вычисленный по соответствующим ЛП-коэффициентам, теоретически мало отличался бы от исходного голосового источника. Для определения интервала закрытых голосовых складок в двухступенчатой ЛП-схеме – 2 используется подвижное окно, в котором вычисляется энергия сигнала-остатка. Предполагается, что на закрытой фазе голосовой щели значение энергии будет минимальным [2, 6]. Однако на использованном тестовом материале это предположение выполнялось далеко не всегда. Примерно в 40% случаев минимум энергии достигался на открытой фазе голосовых складок, при этом соответствующие значения формантных частот и ширин значительно отличались от исходных значений. Как следствие, точность оценки голосового источника была невысокой. Отсюда был сделан вывод, что минимум энергии сигнала-остатка линейного предсказания не может служить надежным критерием определения фазы сомкнутых голосовых складок. Такой же вывод был сделан авторами работы [3] на материале синхронных измерений речевого сигнала и площадей поперечного сечения голосовой щели. Вторым существенным недостатком двухступенчатого алгоритма ЛП– 2 оказалась его неустойчивость к вариациям длительности окна анализа и его расположения внутри периода ОТ. Неустойчивость проявлялась не только в скачкообразных изменениях формантных частот и ширин, но и в потерях полюсов. Отсюда был сделан вывод о непригодности двухступенчатой схемы ЛП– 2 для решения задачи автоматической обратной фильтрации.

Явление потери полюсов также иногда наблюдалось и в двухступенчатом алгоритме ЛП– 3. Подобные потери могут быть отслежены и скорректированы в интерактивном режиме, однако попытка их корректировки в автоматическом режиме наталкивается на большие трудности. Поэтому данный алгоритм также был отвергнут.

Алгоритмы адаптивного определения предсказывающего фильтра, кепстрального разложения и нечеткой кластеризации дали весьма близкие результаты (ошибки порядка 10 – 15 %). В большинстве случаев результаты обратной фильтрации, полученные с помощью этих алгоритмов, оказались хуже результатов фильтрации с помощью асинхронных схем (автокорреляционного и ковариационного алгоритмов). Во всех случаях оценка голосового источника была значительно хуже, чем оценка источника с помощью двухступенчатого линейного предсказания ЛП–1. Дополнительным недостатком алгоритма кепстрального разложения оказалась его крайняя чувствительность к ошибкам в оценке периода ОТ и к расположению временного окна анализа.

ДПМ-алгоритм был предложен в работе [43] как альтернатива методу линейного предсказания. Предварительные эксперименты с синтетическими гласными показали, что ошибки определения формантных частот с помощью этого алгоритма оказались в среднем на 5 % ниже, чем соответствующие ошибки для стандартных ЛП-схем, асинхронных с ОТ. С другой стороны, точность определения голосового источника с помощью обоих ДПМ-алгоритмов была в среднем чуть выше точности обратной фильтрации с помощью асинхронных ЛП-алгоритмов и зачастую гораздо ниже точности двухступенчатой схемы ЛП–1. Попытка запрограммировать синхронный с ОТ ДПМ-алгоритм закончилась неудачей – ошибки оценки голосового источника сразу увеличились на 5 – 10 %. Необходимо отметить, что, в отличие от стандартного линейного предсказания, в ДПМ-алгоритме коэффициенты модели определяются из системы нелинейных уравнений, которая решается итеративно. Это обстоятельство приводит к весьма высоким вычислительным затратам. Кроме того, не исключено возникновение ситуации неустойчивости, вызванной накоплением ошибок вычислений. Поэтому в задаче обратной фильтрации ДПМ-алгоритм не имеет несомненного преимущества перед линейным предсказанием, асинхронным с ОТ.

Таким образом, на материале синтетических гласных наиболее эффективными (как с точки зрения точности обратной фильтрации, так и с точки зрения вычислительных затрат) оказались три схемы – две ЛП – схемы, асинхронные с периодом основного тона, и двухступенчатый ЛП – алгоритм – 1. Эти схемы были апробированы на материале реальных произнесений, записанных синхронно с функциями площадей голосовой щели. Вычисления показали, что обе асинхронные схемы дали физически правдоподобные (и очень близкие) результаты, в то время как двухступенчатая схема – 1 в двух случаях пропустила высокочастотный полюс. Это привело к физически неправильной форме для голосового импульса. Отсюда был сделан вывод о том, что асинхронные ЛП–схемы обратной фильтрации оказываются наиболее предпочтительными методами в полностью автоматических методах оценки параметров голосового источника.

### 3. Тестирование

Тестирование алгоритмов решения обратных задач выполнялось на материале гласных, синтезированных по заранее заданным параметрам голосовой щели, а также для реальных синхронных измерений речевого сигнала и функции площади голосовой щели. В качестве тестовых синтетических гласных были выбраны кардинальные гласные /a/, /i/, /u/ с формантными частотами, указанными в Табл. 2.



Табл. 2. Формантные частоты для синтетических гласных /a/, /i/, /u/

	/a/	/i/	/u/
$F_1$ (Гц)	750	295	321
$F_2$ (Гц)	1257	2159	837
$F_3$ (Гц)	2370	2792	2246
$F_4$ (Гц)	2866	3500	3233
$F_5$ (Гц)	4600	4500	4374

Для всех гласных ширина первых трех формант полагалась равной 100 Гц, а ширины четвертой и пятой формант – 120 Гц и 150 Гц, соответственно. Все три гласных звука были синтезированы для трех различных частот основного тона ( $F_0 = 80$  Гц, 100 Гц, 200 Гц) и четырех различных конфигураций голосовой щели, параметры которых определены в Табл. 3. Типы фонации, соответствующие той или иной конфигурации голосовой щели, указаны в крайнем левом столбце [44, 45]. Данные о конфигурациях голосовой щели были заимствованы из работ [6, 8]. Для всех гласных  $\alpha = 2.0$ ,  $\beta = 1.0$ .

Табл. 3. Параметры голосовой щели, использованные в синтезе.

Тип голоса	Длительность фазы открытия голосовой щели относительно периода основного тона	Длительность фазы открытой голосовой щели относительно периода основного тона
Нейтральный голос	0.41	0.58
Фальцет	0.48	0.72
Придыхательный голос	0.46	0.77
Скрипучий голос	0.3	0.5

Таким образом, общее число синтезированных гласных было равно 36 (3 гласных  $\times$  3 частоты основного тона  $\times$  4 конфигурации голосовой щели).

Синтез гласных осуществлялся с помощью каскадного формантного синтезатора, работающего на частоте дискретизации 10 кГц с 16-битным квантованием каждого отсчета синтетического сигнала. Длительность каждого синтезированного гласного была равна 30-ти периодам основного тона. Моделирование случайных вариаций формантных частот и затуханий, а также периодов основного тона и амплитуд объемных скоростей не проводилось. Подробности алгоритма синтеза описаны в [46].

Помимо синтетических гласных были использованы синхронные измерения реального речевого сигнала и соответствующих площадей голосовой щели. Данные представляли собой несколько периодов гласного /a/ в произнесении четырех женщин (с частотой основного тона от 150 Гц до 325 Гц), а также синхронные измерения площади голосовой щели, выполненные с помощью скоростной киносъемки [47].

#### 4. Вариационный метод

После режектирования полюсов передаточной функции речевого тракта в речевом сигнале в идеале должна остаться функция, пропорциональная производной от объемной скорости через голосовую щель. Поэтому первый метод решения обратной задачи состоял в оценке параметров площади голосовой щели путем согласования вычисленной по уравнениям (4), (6), (7) производной  $w'$  с сигналом, оставшимся после обратной фильтрации.

Параметры функции площади голосовой щели находились путем минимизации функционала Тихонова:

$$\Phi(z_s) = \inf \left\{ \|u_w - A(z)\|_n^2 + \gamma \frac{\|z - z_0\|_n^2}{\|z_0\|_n^2}, z \in Z \right\} \quad (9)$$

Здесь  $Z$  – множество всех физиологически допустимых наборов параметров модели площади голосовой щели;  $z_0$  – начальные приближения,  $A(z)$  – оператор, определяющий по элементу множества  $Z$  импульс голосового источника  $w'$ ,  $\gamma$  – параметр регуляризации (порядка  $1.0e^{-3}$ ),  $u_w$  – некоторый вектор акустических параметров, оцененных по результатам обратной фильтрации.

Пространство  $Z$  параметров математической модели было определено как пятимерное евклидово пространство. В качестве компонент пространства использовалось 4 параметра математической модели голосовой щели (7), а также некоторая константа. Эта константа моделирует постоянную составляющую речевого сигнала, которую невозможно определить с помощью обратной фильтрации без специальной маски Розенберга. Ограничения на множество параметров модели площади голосовой щели были подобраны вручную путем перебора различных значений этих параметров, отображения соответствующих функций площади голосовой щели и объемной скорости на графике и визуальной оценки степени физического и физиологического правдоподобия этих функций. Что касается постоянной составляющей потока, то ее величина по модулю была ограничена единицей, поскольку все значения голосового источника, вычисленного по модели площади голосовой щели (7), перед оптимизацией относились к максимуму модуля этого источника на текущем периоде основного тона.

В качестве  $z_0$  для оптимизации использовались векторы из кодовой книги векторов площади голосовой щели. Две кодовые книги (одна для мужских голосов и одна для женских) были построены путем решения обратной задачи относительно площади голосовой щели и сохранения результатов для 10 дикторов из обучающей выборки (5 мужчин и 5 женщин). Схема построения этих кодовых книг аналогична схеме, описанной в [48]. Объем кодовой книги для мужских голосов составил (после векторного квантования) 256 векторов, а для женщин – 512 векторов.

В качестве процедуры оптимизации функционала Тихонова был использован алгоритм, основанный на последовательной квадратичной аппроксимации функции Лагранжа и решении соответствующей задачи квадратичного программирования квазиньютоновским методом. Этот алгоритм реализован в стандартном модуле `fmincon` пакета MATLAB (Optimization Toolbox).

Для  $u_w$ , совпадающего с экспериментально определенной функцией  $w'_{эксн}$ , результаты такой аппроксимации оказались неудовлетворительными, поскольку ошибки достигали 30 – 40%.

Затем была исследована модель, описывающая не саму площадь голосовой щели, а ее производную. Эта модель первой производной от площади голосовой щели была заимствована из работы [22]:

$$\frac{dS(t)}{dt} = \begin{cases} A_1 \sin\left(\frac{\pi t}{2T_1}\right), & 0 \leq t \leq T_1 \\ (A_1 + A_2) \cos\left(\frac{\pi(t - T_1)}{2(T_2 - T_1)}\right) - A_2, & T_1 < t \leq T_2 \\ -A_2 \left(\frac{T_3 - t}{T_3 - T_2}\right)^2, & T_2 < t \leq T_3 \\ 0, & T_3 < t \leq T_0 \end{cases} \quad (10)$$

Здесь  $T_1$  – момент времени, соответствующий максимальному значению  $A_1$  производной от площади;  $T_2$  – момент времени, соответствующий негативному пику  $A_2$  производной;  $T_3$  – момент времени, соответствующий третьему пересечению производной нулевого уровня (Рис. 2). Соответствующее евклидово пространство  $Z$  также было пятимерным, где первые четыре компонента соответствовали параметрам  $(T_1, T_2, T_3, A_2)$ , а пятым параметром была постоянная составляющая голосового источника.

Искомая функция площади голосовой щели определяется путем численного интегрирования выражения (10). Для того, чтобы избежать проблемы определения постоянной составляющей после интегрирования, было принято, что площадь под функцией производной равна нулю:

$$\int_0^{T_0} \frac{dS_g(t)}{dt} dt = 0$$

Это позволило определить параметр  $A_1$ , тогда как в исходной модели в [22] этот параметр полагался независимым. Затем, как и в первом случае, объемная скорость потока через голосовую щель и ее производная вычислялись по уравнениям (4) и (6).

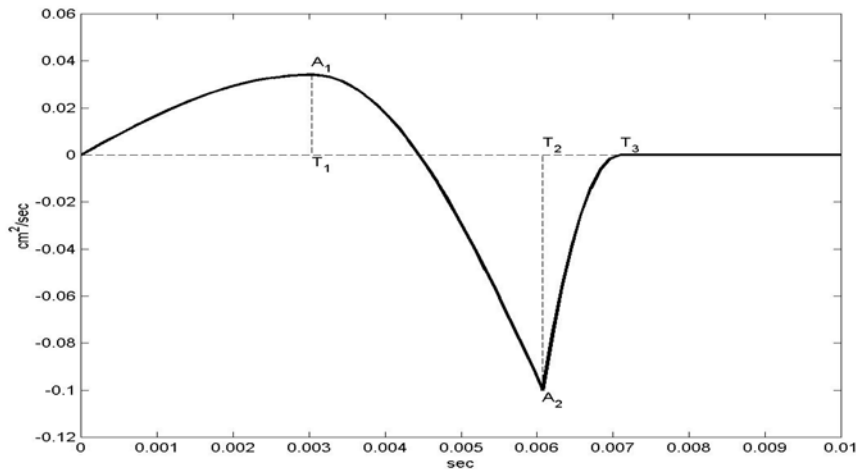


Рис. 2. Модель производной от площади голосовой щели.

Оказалось, что и в этом случае минимизация невязки между вычисленной и экспериментально определенной производной от объемной скорости приводит к слишком большим погрешностям вследствие принципиальной неточности выполнения обратной фильтрации. Поэтому на функции  $w'_{\text{эксн}}$  были определены характерные моменты времени, которые казались наиболее стабильным для разных людей, в разных контекстах и типах микрофонов.

Сначала по функции  $w'_{\text{эксн}}$  определялись положения отрицательных пиков, совпадавших с моментами максимального возбуждения речевого тракта. Интервал между двумя соседними отрицательными пиками  $t_{p1}$  и  $t_{p2}$  принимался за текущий период основного тона. Затем на текущем периоде определялись  $t_{\text{max}}$  – момент времени, соответствовавший максимальному значению  $w'$  на текущем периоде, и  $t_{\text{th}}$  – момент времени, который соответствовал значению  $w'$ , равному 40% от максимального значения источника  $w'_{\text{эксн}}$ ,  $t_{p1} < t_{\text{th}} < t_{\text{max}}$  (Рис. 3). Моменты времени ( $t_{\text{th}}$ ,  $t_{\text{max}}$ ,  $t_{p2}$ ) и значения функции  $w'$  на них и служили шестимерным вектором  $u_w$  параметров голосового источника на текущем периоде основного тона.

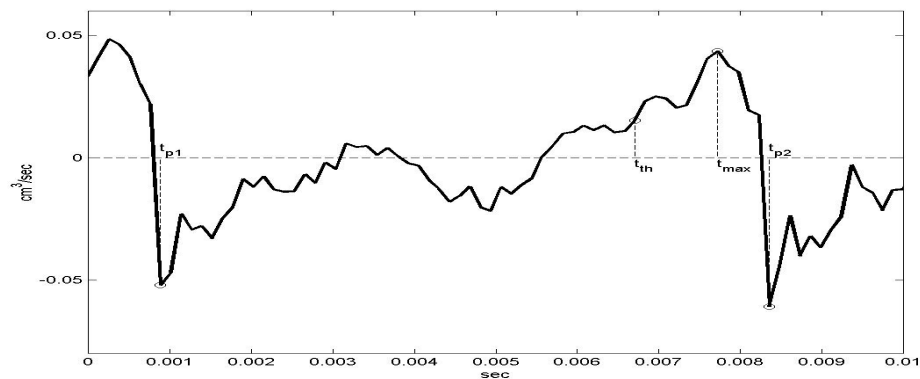


Рис. 3. Характерные точки сигнала после режектирования полюсов.

Все измерения обезразмеривались следующим образом: значения голосового источника относились к максимуму модуля голосового источника на текущем периоде основного тона, а времена – к текущему периоду ОТ.

Предварительные эксперименты показали, что для вектора характерных времен  $u_w$  модель (10) имеет существенные преимущества перед моделью (7). Поэтому эта модель и использовалась в процедуре оптимизации.

Для синтетических гласных средняя ошибка аппроксимации площади составила около 15%. Для синхронных измерений акустического сигнала и площадей голосовой щели ошибка оказалась чуть выше

– порядка 17%. Ошибки в оценке площади росли с увеличением частоты основного тона и понижением частоты первой форманты.

После тестирования на синтезированных звуках этот алгоритм был апробирован на базе данных для 90 дикторов (50 мужчин и 40 женщин), многократно произносивших числительные русского языка от 0 до 9. Ввод речевых сигналов осуществлялся через 4 типа микрофонов: микрофон с шумоподавлением, расположенный вблизи рта диктора; всенаправленный микрофон, расположенный на груди диктора; направленный микрофон, расположенный на расстоянии 40 - 70 см от диктора, и кардиоидный микрофон, расположенный на расстоянии 60 – 90 см от диктора. Помимо различных амплитудно-частотных характеристик приемников звука, их разное расстояние от рта диктора позволяет оценить роль реверберации помещений и влияние как стационарных, так и нестационарных шумов.

На этой базе данных для реальных речевых сигналов вскрылись новые проблемы этого алгоритма. Функции  $w'_{эксн}$ , оцененные с помощью обратной фильтрации, для ряда дикторов и ряда произнесений обнаруживали случайные отрицательные всплески, по амплитуде значительно превосходящие отрицательные пики голосового источника (Рис. 4). Наличие таких пиков зачастую приводило к грубым ошибкам в оценке текущего периода основного тона и соответствующих характерных времен. Примерно для половины дикторов характерные времена оценивались с большой дисперсией.

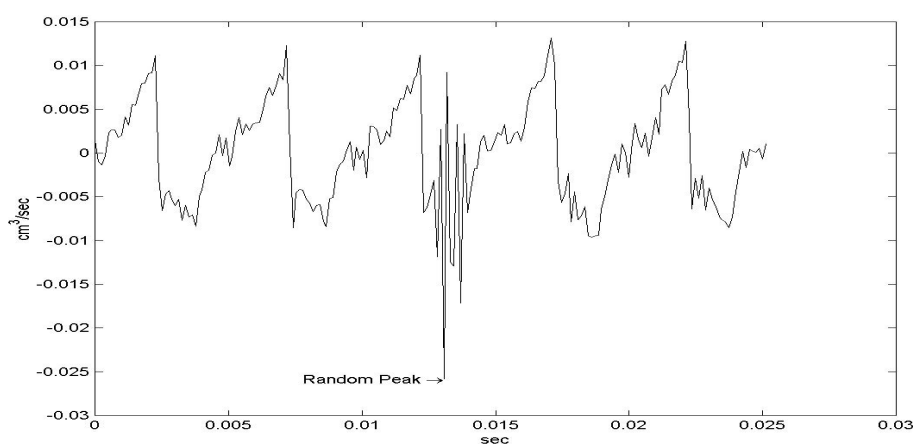


Рис. 4. Неполное режектирование полюсов.

Разброс оценок параметров функции площади голосовой щели, конечно, определяется, в первую очередь, неточностью модели линейного предсказания, не учитывающей влияние акустического канала и характеристик микрофона. В результате, найденная путем обратной фильтрации функция голосового источника  $w'_{эксн}$  может быть искажена настолько, что поиск наилучшей функции площади голосовой щели по схеме интегрирования (10) с последующим вычислением потока и его производной с использованием регуляризации (9) не обеспечивает характеристик, необходимых для практического применения, например, для верификации диктора по его голосовому источнику. Это может рассматриваться как указание на неприемлемость использования производной от объемной скорости в качестве единственного источника информации для вычисления невязки в (9).

## 5. Среднеквадратическая аппроксимация.

Нестабильность оценок площади голосовой щели путем непосредственного сравнения вычисленной и измеренной производной от объемной скорости привела к исследованию алгоритма решения обратной задачи для голосового источника, основанного на использовании уравнения (8). Если по речевому сигналу каким-то образом измерена объемная скорость потока  $w(t)$ , то можно попытаться вычислить площадь голосовой щели  $S(t)$ , варьируя параметры  $b$ ,  $h$  и  $\Delta p$  и применяя критерий минимума невязки между измеренной и вычисленной объемной скоростью. Невязка может вычисляться и относительно вычисленной и измеренной площади голосовой щели. Фактически, это обратная задача со всеми вытекающими отсюда последствиями – неустойчивостью и неоднозначностью решения. Однако, вопреки опасениям, оказалось, что достаточно точные решения могут быть получены и без привлечения техники решения обратных задач.

Наиболее удачным оказался алгоритм, в котором сигнал, полученный после обратной фильтрации, интегрировался и вычисленные таким образом оценки объемной скорости служили входными данными для определения  $S_{эксн}$  по уравнению (8). Функция  $S_{эксн}$  затем аппроксимировалась по критерию минимума среднеквадратической ошибки функцией (7). Некоторые математические свойства такой задачи обеспе-

чивают единственное устойчивое решение, которое получается просто минимизацией невязки, без учета стабилизирующего функционала [49]. Этот алгоритм удачно прошел тестирование на синтезированных сигналах с известной площадью голосовой щели, а также на сигналах с измеренной площадью голосовой щели.

На Рис. 5 показаны осциллограмма синтезированного звука, истинная функция площади голосовой щели, использованная для создания голосового источника с помощью уравнения (2), и решение уравнения (8) после обратной фильтрации голосового источника. Установлено, что, несмотря на принятые допущения, исходная и вычисленная площадь голосовой щели для синтезированных сигналов практически совпадают. При этом погрешность аппроксимации исходной площади моделью (7) составляет доли процента.

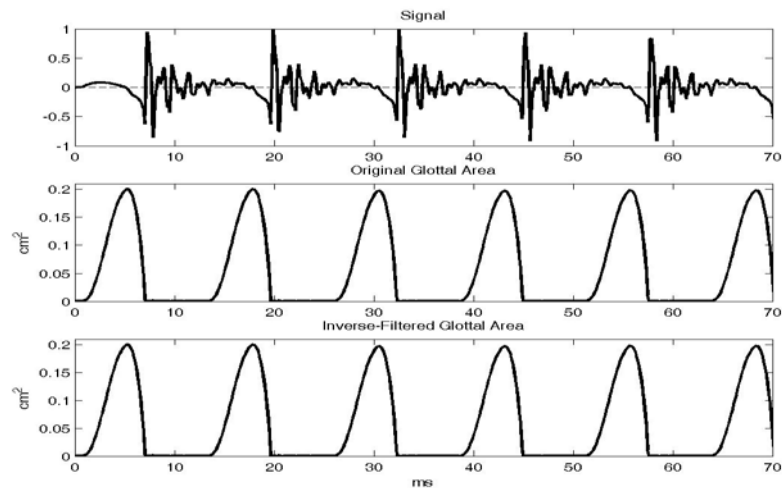


Рис. 5. Площадь голосовой щели в синтезированном сигнале и ее оценка путем решения уравнения (8) после обратной фильтрации.

Следующий этап тестирования состоял в вычислении  $S(t)$  для речевых сигналов, которые были записаны синхронно со сверхскоростной регистрацией площади голосовой щели диктора. На Рис. 6, показаны измеренные и вычисленные функции  $S(t)$  для разных значений основного тона. Видно, что и в этом случае измеренная и вычисленная функция площади достаточно близки. Ошибки аппроксимации для сигналов с синхронным измерением площади голосовой щели моделью (7) составили, в среднем, 10 – 12 %. При этом следует оговориться, что измерения площади голосовой щели были представлены в некоторых абстрактных единицах, а абсолютные значения остались неизвестными. Поэтому о степени точности полученного решения можно судить лишь по совпадению формы импульсов, тогда как максимальные значения измеренной площади и точность ее вычисления все же неизвестны. К тому же, техника решения обратной задачи не позволяет вычислить постоянный уровень площади голосовой щели, который часто наблюдается у женских голосов вследствие неполного закрытия щели. Эта разница видна на нижнем графике Рис. 6.

Апробация алгоритма на базе реальных данных показала следующее.

При интегрировании сигнала-остатка после режектирования полюсов передаточной функции речевого сигнала возникает некоторый тренд, т.е. медленно меняющаяся аддитивная функция. Именно тот факт, что эта функция изменяется во времени гораздо медленнее, чем импульсы объемной скорости, позволяет выполнить фильтрацию тренда нуль-фазовым фильтром высоких частот с частотой среза примерно 30 – 40 Гц [4, 12]. После ликвидации тренда минимальные значения импульсов объемной скорости оказываются ненулевыми и разными по значению. Для большинства режимов голосообразования существует интервал сомкнутых голосовых складок, на котором аэродинамический поток отсутствует и, следовательно, объемная скорость потока также должна быть равна нулю. В некоторых режимах и у некоторых женщин полного смыкания голосовых складок не происходит, так что минимальное значение объемной скорости не равно нулю. Однако эту "постоянную" составляющую потока невозможно вычислить методом обратной фильтрации. Для этого используется специальный прибор, называемый маской Розенберга [50]. Поэтому в нашем алгоритме было принято требование, чтобы минимальное значение объемной скорости равнялось нулю. Это допущение, очевидно, приводит к погрешности оценки функции площади голосовой щели в некоторых случаях, но избежать этого не удастся.

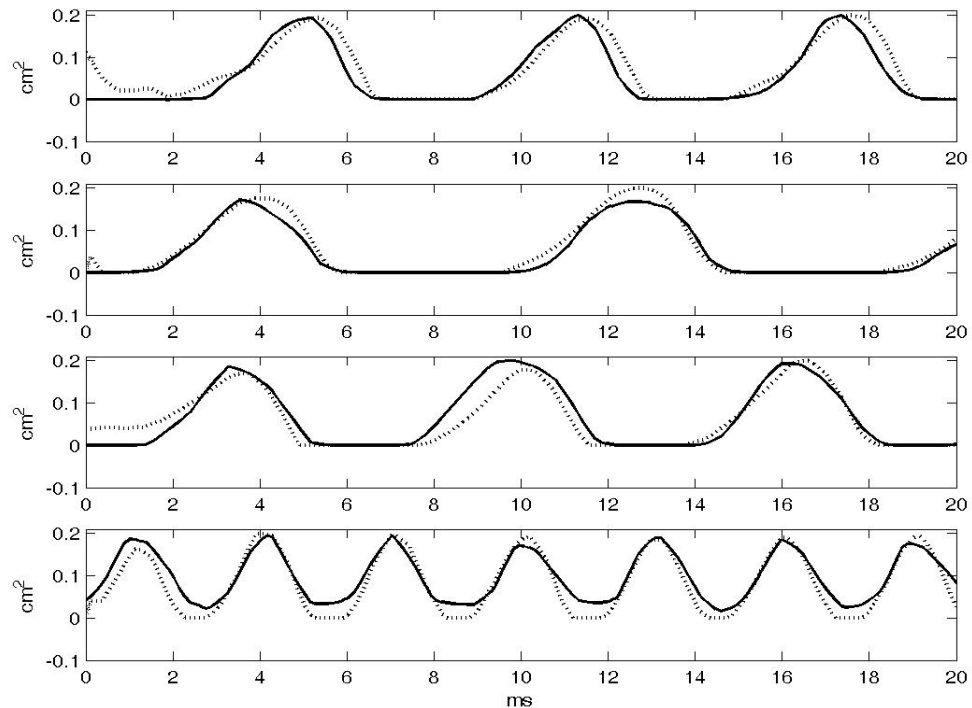


Рис. 6. Измеренные (—) и вычисленные (---) площади голосовой щели.

Вследствие того, что минимальные значения объемной скорости, вычисленные путем интегрирования сигнала-остатка, не одинаковы для одной и той же последовательности импульсов голосового источника, нельзя использовать наименьшее значение для одного из этих минимумов. В противном случае минимальные значения объемной скорости для соседних импульсов голосового источника все равно будут ненулевыми.

Исходя из наблюдений за поведением экспериментально определенных импульсов объемной скорости для синтезированных сигналов с известными интервалами сомкнутых голосовых складок, был сформулирован алгоритм поиска нулевой линии. Согласно этому алгоритму, к каждому минимальному значению объемной скорости между соседними максимумами добавлялась величина, равная некоторой доле от разности между одним из максимумов и этим минимумом,  $\Delta = \delta(w_{\max} - w_{\min})$ . Значение  $w_{\min} + \Delta$  принималось за нулевую линию на интервале от начала импульса, находящегося слева от минимума и до начала импульса, находящегося справа от этого минимума (Рис. 7).

После определения нулевой линии ожидаемый интервал сомкнутых голосовых складок может оказаться слишком малым, а форма импульса объемной скорости вблизи нуля подвержена искажениям вследствие неточности обратной фильтрации. Фильтрация этих искажений представляет собой сложную, трудно решаемую задачу. Поэтому аппроксимация экспериментально определенной по (8) площади голосовой щели  $S_{\text{эсп}}$  моделью голосовой щели (7) выполнялась не на всем интервале ненулевых значений  $S_{\text{эсп}}$ , а лишь на том интервале, на котором функция  $S_{\text{эсп}}$  была наименее искажена ошибками обратной фильтрации. Этот интервал определялся с помощью следующего алгоритма.

Пусть  $[n_1 \ n_2]$  – номера начального и конечного отсчетов на интервале текущего импульса объемной скорости (Рис. 7),  $i = 1$  ( $i$  – номер текущей итерации),  $N = 10$  ( $N$  – максимальное число итераций). Тогда выполняется следующая последовательность операций:

1. Выполнить среднеквадратическую аппроксимацию функции  $S_{\text{эсп}}$  моделью (7) на интервале  $[n_1 + i - 1 \ n_2]$ . Вычислить среднеквадратическую ошибку аппроксимации  $E_i$ .
2. Если  $i < N$ , то положить  $i = i + 1$  и перейти к шагу 1. В противном случае перейти к шагу 3.
3. Определить  $m = \text{argmin}(E_i)$ ,  $i = 1, \dots, N-1$ .
4. В качестве искомого интервала выбрать интервал  $I = [n_1 + m - 1 \ n_2]$ .

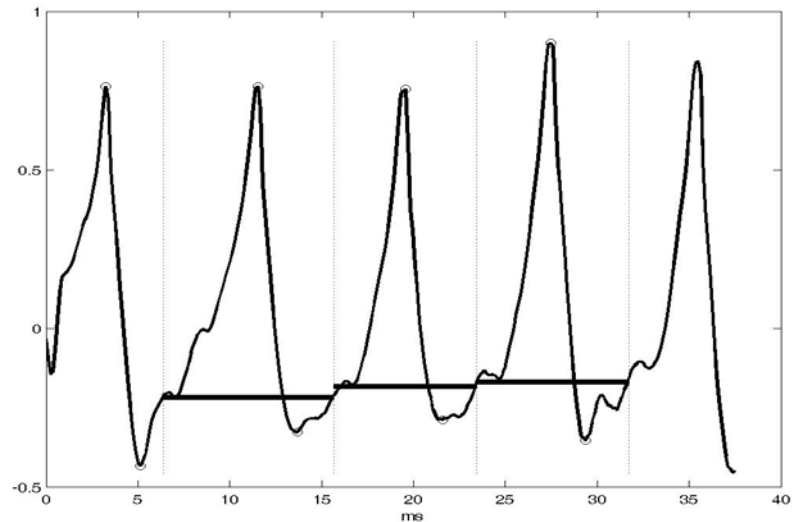


Рис. 7. Определение "нулевой" линии.

Действие этого алгоритма на сигналах, полученных от кардиоидного микрофона, показано на Рис. 8.

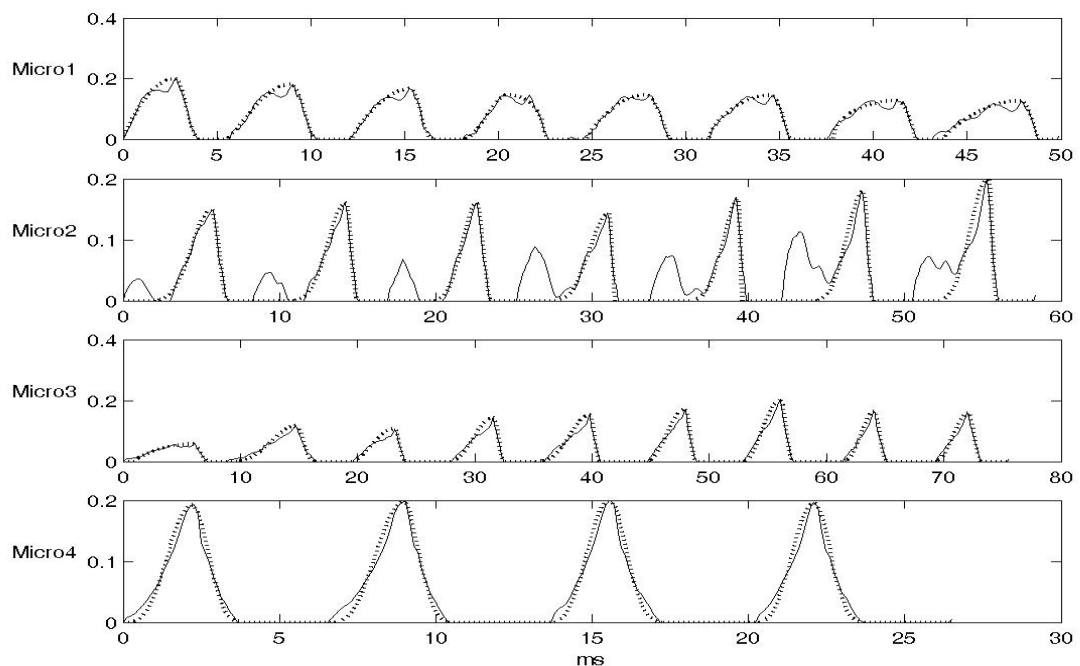


Рис.8. Решение обратной задачи для сигналов, полученных с разных микрофонов. Micro1 – направленный микрофон, Micro2 – кардиоидный микрофон, Micro3 – всенаправленный, Micro4 – головная гарнитура №1.

Для сигналов из базы данных ошибки между площадью голосовой щели  $S_{\text{экс}}(t)$  и функцией  $S(t)$ , порождаемой моделью (7), составили в среднем 3 – 5 %. Эти оценки относятся лишь к точности аппроксимации вычисленной по речевому сигналу площади голосовой щели посредством параметрической модели (7). Однако, поскольку истинная площадь голосовой щели в этих экспериментах была неизвестна, то и точность ее восстановления остается неопределенной. Косвенная оценка точности восстановления площади голосовой щели могла бы быть получена путем сопоставления результатов анализа одного и того же речевого сигнала, записанного одновременно через разные микрофоны.

В нашей базе данных нашелся такой диктор, речь которого была записана через три пары различных микрофонов. Это давало надежду на дополнительную оценку устойчивости анализа в предположении, что площадь голосовой щели для данного диктора в разных произнесениях должна была бы быть похожей.

В первой серии экспериментов использовались записи речи через телефонную трубку и направленный микрофон, укрепленный на груди диктора. Во второй серии использовалась телефонная трубка другого размера и кардиоидный (с полусферической диаграммой направленности) микрофон, установленный на мониторе компьютера на расстоянии примерно 60 – 90 см от диктора. В третьей серии экспериментов использовалась головная гарнитура с шумоподавляющим микрофоном и кардиоидный микрофон, установленный аналогично второй серии экспериментов. Анализировались сегменты ударных гласных в словах */два, три, восемь/*.

Такие типы и расположение микрофонов позволяют оценить влияние различных неречевых факторов. Микрофоны, близко расположенные ко рту диктора, характеризуются подъемом уровня низких частот из-за свойств ближнего акустического поля. Наряду с подъемом низких частот, телефонные трубки создают также и подъем высоких частот вследствие влияния небольшой воздушной полости, в которой находится микрофон, причем степень этого подъема зависит от объема этой полости, различного у разных трубок. Удаленный микрофон не искажает амплитудно-частотную характеристику речевого сигнала, однако воспринимает реверберационный отклик помещения.

На Рис. 9, 10 и 11 показаны результаты вычисления площади голосовой щели по (8) для каждой пары микрофонов.

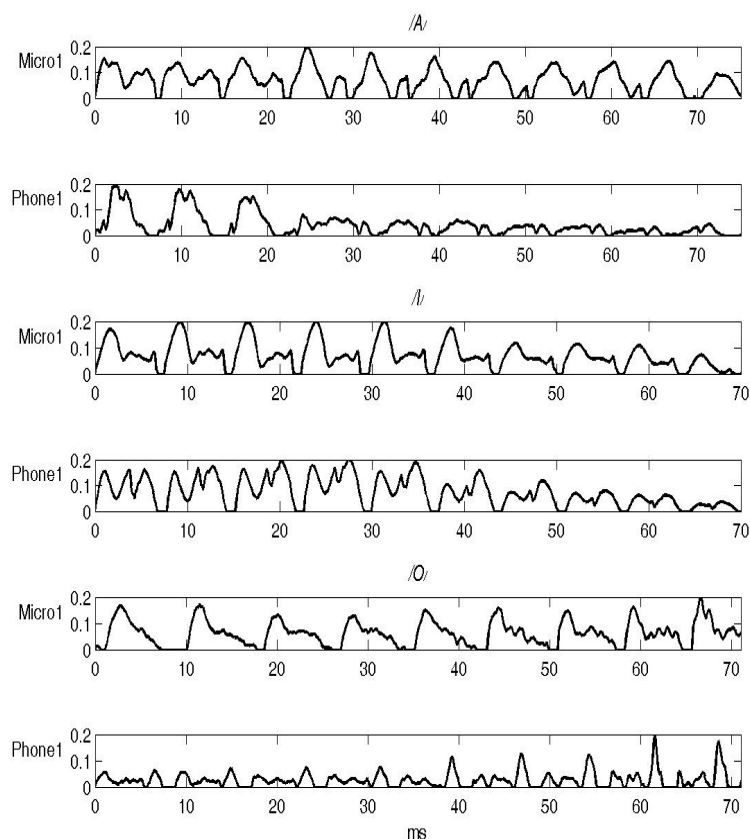


Рис. 9. Вычисление площади голосовой щели непосредственно после обратной фильтрации. Micro1 – направленный микрофон на груди диктора, Phone1 – телефонная трубка №1.



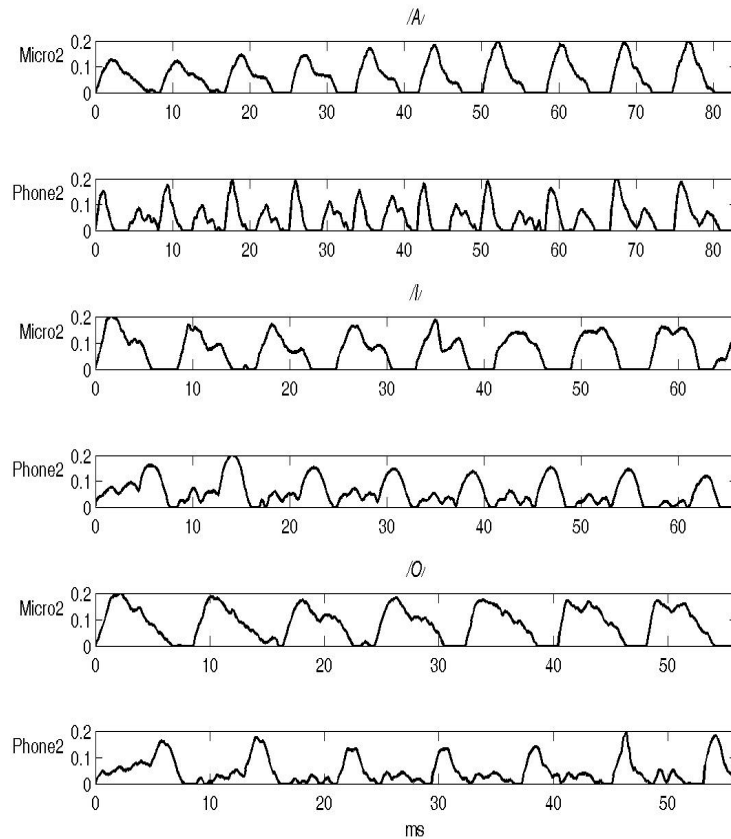


Рис. 10. Вычисление площади голосовой щели непосредственно после обратной фильтрации. Micro2 – кардиоидный микрофон, Phone2 – телефонная трубка №2.

В стандартном методе линейного предсказания искажения, вносимые приемником звука, сравнительно мало влияют на восприятие сигнала, восстановленного на приемном конце канала связи. В обратной задаче для голосового источника это влияние очень велико. Сопоставление результатов анализа для каждой пары микрофонов приводит к выводу, что тип и расположение микрофона существенно влияют на вычисление площади голосовой щели и, следовательно, на параметры модели (7), используемой для аппроксимации экспериментальных данных. Поэтому либо необходимо приспосабливать метод линейного предсказания к характеристикам приемника звука, либо использовать только такие приемники, которые мало искажают площадь голосовой щели.

В наших экспериментах было обнаружено, что далеко расположенные микрофоны иногда приводят к заметному искажению речевого сигнала. Для уточнения этого явления был проведен специальный эксперимент с целью оценки влияния реверберации помещения. Синтезированные гласные /a, i, u/ по 10 раз воспроизводились через динамик и принимались направленным микрофоном высокого качества, расположенным на расстоянии примерно 40 см от динамика. Поскольку площадь голосовой щели в синтезированном сигнале точно известна, и воспроизводился один и тот же сигнал, то можно было бы оценить расхождение между истинной и вычисленной площадью, учитывая задержку в распространении сигнала от динамика к микрофону около 1 мс. Ожидалось, что такое расхождение будет невелико.

Результаты этого эксперимента оказались драматическими. Вычисленные площади голосовой щели, во-первых, мало походили на исходную площадь и, во-вторых, сильно отличались не только при разных воспроизведениях гласного, но и внутри одной и той же реализации в последовательности импульсов голосового возбуждения. На Рис. 12 показана последовательность истинных функций площади голосовой щели для трех импульсов и наложение вычисленных площадей по 10 реализациям.

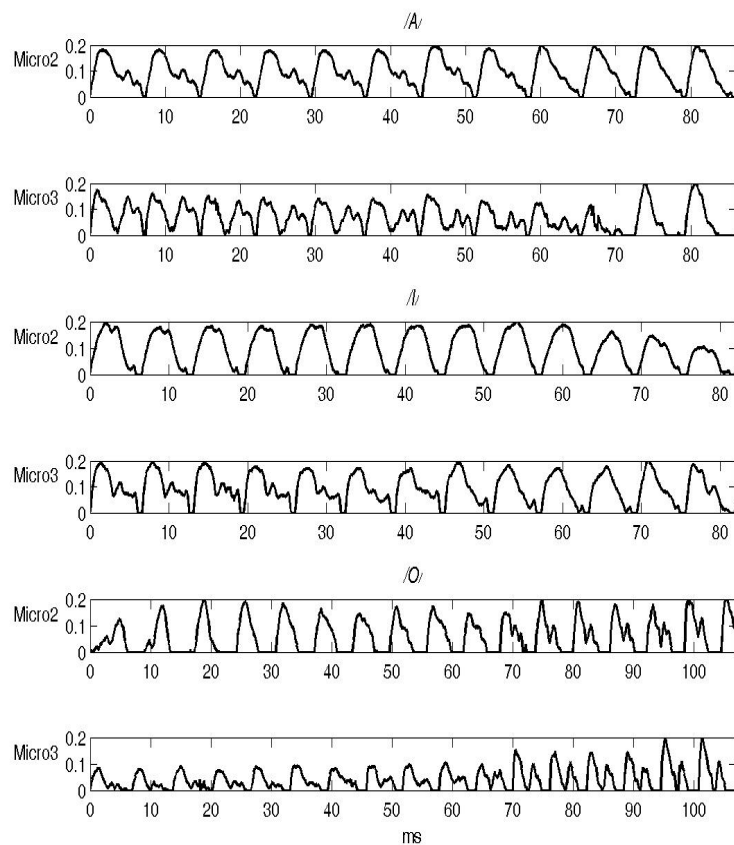


Рис. 11. Вычисление площади голосовой щели непосредственно после обратной фильтрации. Micro2 – кардиоидный микрофон, Micro3 – головная гарнитура №2.

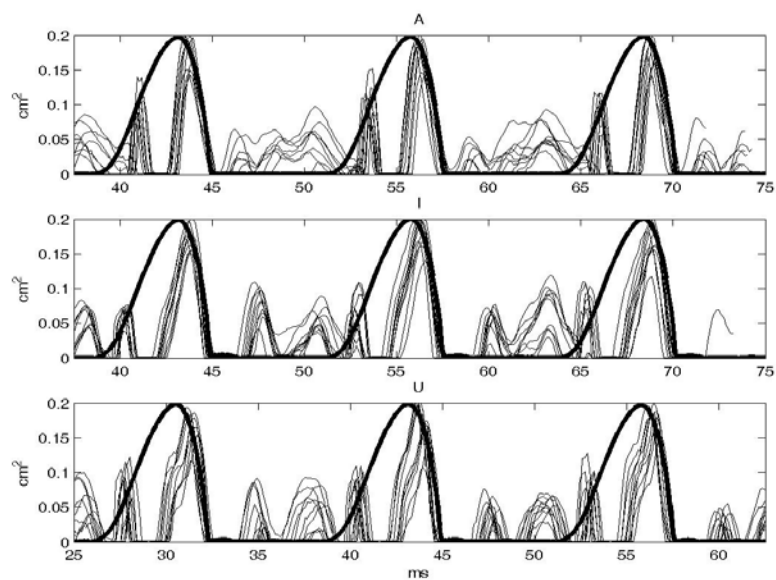


Рис. 12. Площадь голосовой щели в синтезированном сигнале (жирная линия) и ее оценки в эксперименте с воспроизведением сигнала через динамик.

С целью выяснения причин такого поведения был выполнен эксперимент по оценке характеристик взаимодействия системы "динамик - микрофон". Как известно, белый шум имеет равномерный по частоте спектр. Поэтому, проигрывая его через динамик, можно оценить вносимые искажения. В нашем эксперименте тестовый сигнал был получен с помощью процедуры, синтезирующей равномерное распределение на интервале от 0 до 1. Как видно из Рис. 13, спектр этого сигнала почти равномерен на интервале до 4 кГц и, что важнее, это монотонная функция. Это дает возможность определить характер искажений в системе "динамик - микрофон". Очевидно, в этой системе происходит существенное искажение спектральных характеристик – появились нули и полюсы. Особенно четко выражен полюс на частоте около 300 Гц и нуль на частоте 1000 Гц. Такая сильная неравномерность амплитудно-частотной характеристики вряд ли является свойством динамика. Эти дополнительные полюсы и особенно нули не могут быть определены канонической процедурой линейного предсказания и, следовательно, скомпенсированы путем обратной фильтрации. Колебания на частоте дополнительного полюса с частотой 300 Гц хорошо видны на Рис. 12.

Таким образом, выяснилась существенная роль акустического канала, хотя разделить влияние реверберации и свойств динамика не удастся.

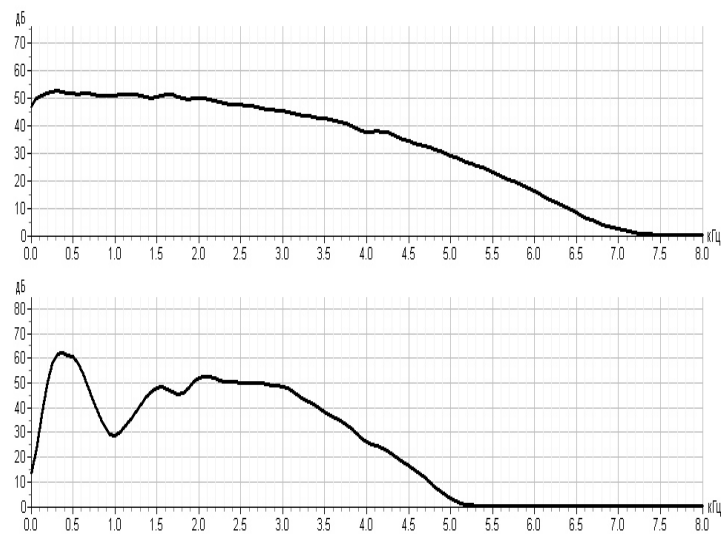


Рис. 13. Спектр тестового сигнала (вверху) и сигнала, пропущенного через систему "динамик - микрофон" (внизу).

Нестабильность оценок площади голосовой щели, очевидная из Рис. 12, указывает на то, что не только искажена амплитудно-частотная характеристика сигнала, но присутствуют и вариации сигнала во времени. Это можно отнести только к реверберации помещения, создающей биения частотных компонент речевого сигнала. Эти биения видны и на Рис. 8 для сигналов, принятых микрофоном, удаленным от диктора на расстояние до 80 см. Отсюда следует вывод, что наименьшая погрешность в оценке площади голосовой щели путем обратной фильтрации и последующем решении обратной задачи достигается только для микрофонов, расположенных достаточно близко к диктору.

## 6. Заключение

Некорректная обратная задача относительно формы функции, описывающей изменения площади голосовой щели во времени, сведена к задаче минимизации невязки между функцией, порождаемой моделью голосового источника  $S_{\text{мод}}(t, t_1, t_2, \alpha, \beta)$ , и функцией  $S_{\text{экл}}(t)$ , вычисленной с использованием метода обратной фильтрации. Решение этой задачи для синтезированных речевых сигналов с известным голосовым источником и для сигналов с измеренной площадью голосовой щели оказалось вполне удовлетворительным, поскольку среднеквадратические ошибки находились в диапазоне до 0.1 % (для синтетических гласных) и 10 – 12 % (измеренные площади). При тестировании этого алгоритма на обширной базе данных для 90 дикторов распределение вычисленных параметров  $(t_1, t_2, \alpha, \beta)$  находилось в физиологически правдоподобных пределах. Это позволяет рассматривать разработанный метод решения обратной задачи подходящим как для исследования свойств голосового источника, так и для применения в речевых технологиях, таких, как синтез речи по тексту, распознавание речи, верификация и идентификация дикторов, а также сжатие речи.

Вместе с тем, выяснилось, что метод линейного предсказания, который используется для обратной фильтрации, неустойчив относительно типа приемника звука. Отсюда следует, что приемлемые оценки функции площади голосовой щели могут быть получены только для направленных микрофонов, расположенных на небольшом расстоянии от диктора.

## Литература

1. R. Miller, "Nature of the Vocal Cord Wave", J. Acoust. Soc. Amer., vol. 31, pp. 667-677, 1959.
2. D. Wong, J. Markel, A. Gray, "Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform", IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-27, No 4, pp. 350-355, 1979.
3. J. Larar, Y. Alsaka, D. Childers, "Variability in Closed Phase Analysis in Speech", Int. Conf. Acoust., Speech, Signal Process., pp. 1089 – 1092, 1985.
4. D. Veeneman, S. BeMent, "Automatic Glottal Inverse Filtering from Speech and Electroglottographic Study", IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-33, No. 2, pp. 369-377, 1985.
5. A. Krishnamurthy, D. Childers, "Two-Channel Speech Analysis", IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-34, No. 4, pp. 730-743, 1986.
6. D. Childers, Ch. Ahn, "Modeling the Glottal Volume Velocity Waveform for Three Voice Types", J. Acoust. Soc. Amer., vol. 97, No. 1, pp. 505 – 519, 1995.
7. M. Matausek, V. Batalov, "A New Approach to the Determination of the Glottal Waveform", IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-28, No. 6, pp. 616-622, 1980.
8. D. Childers, C. Lee, "Vocal Quality Factors: Analysis, Synthesis, and Perception", J. Acoust. Soc. Amer., vol. 90, No. 5, pp. 2394 – 2410, 1991.
9. P. Alku, "Glottal Wave Analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering", Speech Commun., vol. 11, pp. 109-118, 1992.
10. M. Plumpe, T. Quatieri, D. Reynolds, "Modeling the Glottal Flow Derivative with Application to Speaker Identification", IEEE Trans. Speech, Audio Process., vol. 7, No. 5, pp. 569-585, 1999.
11. O. Akande, P. Murphy, "Estimation of the Vocal Tract Transfer Function with Application to Glottal Wave Analysis", Speech Commun., vol. 46, No. 1, pp. 15-36, 2005.
12. P. Milenkovic, "Glottal Inverse Filtering by Joint Estimation of an AR System with a Linear Input Model", IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-34, No. 1, pp. 28-42, 1986.
13. A. Isaksson, M. Millnert, "Inverse Glottal Filtering Using a Parametrized Input Model", Signal Process., vol. 18, No. 4, pp. 435-446, 1989.
14. Q. Fu, P. Murphy, "Robust Glottal Source Estimation Based on Joint Source-Filter Model Optimization", IEEE Trans. Audio, Speech, Language Process., vol. 14, No. 2, pp. 492-501, 2006.
15. M. Froelich, D. Michaelis, H. W. Strube, "SIM – Simultaneous Inverse Filtering and Matching of a Glottal Flow Model for Acoustic Speech Signals", J. Acoust. Soc. Amer., vol. 110, No. 1, pp. 479-488, 2001.
16. T. Backstrom, P. Alku, E. Vilkman, "Time-Domain Parametrization of the Closing Phase of Glottal Air-flow Waveform from Voices over a Large Intensity Range", IEEE Trans. Speech, Audio Process., vol. 10, No. 3, pp. 186-192, 2002.
17. A. Alkhairy, "An Algorithm for Glottal Volume Velocity Estimation", Proc. Int. Conf. Acoust., Speech, Signal Process., 1999.
18. Y. Shapira, I. Gath, "A Geometrical Fuzzy Clustering-Based Solution to Glottal Wave Estimation", J. Acoust. Soc. Amer., vol. 104, No. 5, pp. 3070-3079, 1998.
19. А. Н. Тихонов, А. С. Леонов, А. Г. Ягола. *Нелинейные некорректные задачи*. – М.: Наука. Физматлит. 1995. – 312 с.
20. В. Н. Сорокин. *Теория речеобразования*. – М.: Радио и Связь. 1985. – 312 с.
21. В. Н. Сорокин. *Синтез речи*. – М.: Наука. 1992. – 392 с.
22. T. Ananthapadmanabha, "Acoustic Analysis of Voice Source Dynamics", STL-QPSR, No. 2-3, pp. 1-24, 1984.
23. H. Strik, L. Boves, "On the Relation between Voice Source Parameters and Prosodic Features in Connected Speech", Speech Commun., No. 11, pp. 167-174, 1992.
24. G. Fant, "Some Problems in Voice Source Analysis", Speech Commun., No. 13, pp. 7-22, 1993.
25. I. Karlsson, J. Liljencrants, "Diverse Voice Qualities: Models and Data", TMH-QPSR, No. 2, pp. 143-146, 1996.
26. G. Fant, "The Voice Source in Connected Speech", Speech Commun., No. 22, pp. 125 – 139, 1997.
27. H. Strik, "Automatic Parametrization of Differentiated Glottal Flow: Comparing Methods by Means of Synthetic Flow Pulses", J. Acoust. Soc. Amer., vol. 103, No. 5, Pt. 1, pp. 2659-2669, 1998.
28. G. Fant, J. Liljencrants, Q. Lin, "A Four Parameter Model of Glottal Flow", STL-QPSR, vol. 4, pp. 1-13, 1985.
29. P. Milenkovic, "Voice Source Model for Continuous Control of Pitch Period", J. Acoust. Soc. Amer., vol. 93, No. 2, pp. 1087-1096, 1993.

30. D. Childers, H. Hu, "Speech Synthesis by Glottal Excited Linear Prediction", J. Acoust. Soc. Amer., vol. 96, No. 4, pp. 2026-2036, 1994.
31. J. Schoentgen, "Glottal Waveform Synthesis with Volterra Shaping Functions", Speech Commun., vol. 11, pp. 499-512, 1992.
32. J. Schoentgen, "Shaping Function Models of the Phonatory Excitation Signal", J. Acoust. Soc. Amer., vol. 114, No. 5, pp. 2906-2912, 2003.
33. K. Narasimhan, J. Principe, D. Childers, "Nonlinear Dynamic Modeling of the Voiced Excitation for Improved Speech Synthesis", Proc. Int. Conf. Acoust., Speech, Signal Process., AZ, pp. 389-392, 1999.
34. E. Rank, G. Kubin, "An Oscillator – Plus – Noise Model for Speech Synthesis", Speech Commun., vol. 48, pp. 775-801, 2006.
35. N. Pinto, D. Childers, A. Lalwani, "Formant Speech Synthesis: Improving Production Quality", IEEE Trans. Acoust., Speech, Signal Process., vol. 37, No. 12, pp. 1870-1887, 1989.
36. S. Gupta, J. Schroeter, "Pitch-Synchronous Frame-by-Frame and Segment-Based Articulatory Analysis by Synthesis", J. Acoust. Soc. Amer., vol. 94, No. 5, pp. 2517-2530, 1993.
37. I. Titze, D. Wong, B. Story, R. Long, "Considerations in Voice Transformation with Physiologic Scaling Principles", Speech Commun., vol. 22, pp. 113-123, 1997.
38. K. Tom, I. Titze, "Vocal Intensity in Falsetto Phonation of a Countertenor: An Analysis by Synthesis Approach", J. Acoust. Soc. Amer., vol. 110, No. 3, pp. 1667-1676, 2001.
39. Q. Lin, "Nonlinear Interaction in Voice Production", STL-QPSR, No. 1, pp. 1 – 12, 1987.
40. I. Titze, S. Mapes, B. Story, "Acoustics of the Tenor High Voice", J. Acoust. Soc. Amer., vol. 95, pp. 1133 – 1142, 1994.
41. Д. Маркел, А. Грей. *Линейное предсказание речи*. - М.: Связь. 1980. – 308 с.
42. W. Chu. *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. A John Wiley and Sons, Inc. P. 578. 2003.
43. A. El-Jaroudi, J. Makhoul, "Discrete All-Pole Modeling", IEEE Trans. Signal Process., vol. 39, No. 2, pp. 411 – 423, 1991.
44. С. В. Кодзасов, О. Ф. Кривнова. *Общая фонетика*. - М., 2001. - 592 с.
45. P. Ladefoged, I. Maddieson. *The Sounds of the World's Languages*. Blackwell Publishing Ltd. P. 426. 1996.
46. А. С. Леонов, И. С. Макаров, В. Н. Сорокин, А. И. Цыплихин, "Артикуляторный ресинтез гласных", Информационные процессы, Т. 3, No. 2, стр. 73-82, 2003.
47. A. Krishnamurthy, D. Childers, "Vocal Fold Vibratory Patterns: Comparison of Film and Inverse Filtering", Proc. Int. Conf. Acoust., Speech, Signal Process., pp. 133-136, 1981.
48. А. С. Леонов, И. С. Макаров, В. Н. Сорокин, А. И. Цыплихин, «Кодовая книга для речевых обратных задач», Информационные процессы, Т. 5, No. 2, стр. 101-119, 2005.
49. А. Н. Тихонов, В. Я. Арсенин. *Методы решения некорректных задач*. – М.: Наука. 1986. – 288 с.
50. M. Rothenberg, "A New Inverse-Filtering Technique for Deriving the Glottal Airflow Waveform During Voicing", J. Acoust. Soc. Amer., vol. 53, No. 6, pp. 1632-1645, 1973.