АКУСТИЧЕСКИЙ ЖУРНАЛ, 2021, том 67, № 2, с. 185–202

ОБРАБОТКА АКУСТИЧЕСКИХ СИГНАЛОВ. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

УДК 612.85

ФАЗОВЫЙ АНАЛИЗ АКТИВНОСТИ ГОЛОСОВОГО ИСТОЧНИКА

© 2021 г. В. Н. Сорокин[a], *, А. С. Леонов[b]

[a] *Институт проблем передачи информации, Российская академия наук, Большой Каретный пер. 19, стр. 1, Москва, 127051 Россия*

[b] *Национальный исследовательский ядерный университет "МИФИ", Каширское ш. 31, Москва, 115409 Россия*

**e-mail: vns@iitp.ru*

Поступила в редакцию 02.05.2020 г.
После доработки 14.12.2020 г.
Принята к публикации 22.12.2020 г.

Предложены математические модели, позволяющие связать параметры голосового источника с параметрами фазово-частотных характеристик (ФЧХ) сегментов речевого сигнала. В частности, установлено, что длительность работы источника можно найти по средней длине интервалов между нулями и точками разрыва этих ФЧХ. Для синтетических и реальных речевых сигналов на основе установленных свойств ФЧХ и предложенных эвристических методов их анализа проведена численная оценка периодов основного тона, длительностей работы голосового источника внутри этих периодов, а также моментов начала T_{op} и конца T_{cl} действия голосового источника. Экспериментально установлено существование верхней границы диапазона частот основного тона F_0, внутри которого ошибка оценки F_0 не превышает 5%. Средняя ошибка оценки длительности голосового источника по предлагаемой методике для сегментов речи из базы данных Arctic оказалась менее 0.3% для двух дикторов, а для третьего диктора равна 6.2%. Показано, что ошибка определения величин T_{op} и T_{cl} зависит от свойств голосового источника и значительно возрастает для $F_0 > 220$ Гц. Наиболее вероятная ошибка оценки величин T_{op} для трех дикторов из базы данных Arctic оценивается как 1.5, 10.2 и 13.5%, а для T_{cl} она составляет –9.7, –20.2 и –13.9%.

Ключевые слова: распознавание речи, идентификация диктора, фазово-частотная характеристика, параметры голосового источника

DOI: 10.31857/S0320791921020088

1. ВВЕДЕНИЕ

Характеристики голосового источника играют заметную роль при идентификации диктора, в распознавании речи и при анализе патологии голоса. Важнейшие из этих характеристик – моменты начала и конца активности голосового источника, которые коррелированны с моментами открытия T_{op} и закрытия T_{cl} голосовой щели, а также частота основного тона F_0. Именно эти параметры необходимо в первую очередь определить при анализе сегмента речи с помощью математического моделирования. Они в значительной мере определяют функциональную форму голосового источника при его описании по известным математическим моделям (см., например, [1] и формулу (7) ниже).

Перед закрытием голосовой щели в речевом сигнале возникает всплеск энергии, который вызывается отрицательным пиком производной от объемной скорости потока через голосовую щель. Связанные с этим пиком явления содержат информацию о моменте закрытия голосовой щели T_{cl}. Для определения величины T_{cl} во временной области часто применяется анализ сигнала-остатка после выполнения анализа методом линейного предсказания (см., например, [1]).

В спектрально-временной области для этих целей используются экстремумы логарифмической производной средней энергии спектра речевого сигнала в области частот второй и третьей форманты [2]. Обширный обзор других методов определения момента закрытия голосовой щели в фазово-частотной области представлен в [3]. Все эти методы позволяют достаточно точно найти моменты закрытия голосовой щели. В то же время, имеющиеся методы определения момента открытия голосовой щели дают значительную погрешность.

В задачах определения формы импульса голосового источника необходимо знать оба этих мо-

мента, и погрешность определения моментов открытия и закрытия голосовой щели существенно влияет на точность восстановления формы импульса. Это продемонстрировано, например, в работе [4], где форма импульсов голосового источника вычисляется путем обратного преобразования Фурье отношения спектров речевого сигнала на интервалах открытой и закрытой голосовой щели.

В исследованиях характеристик голосового источника основное внимание уделяется анализу амплитудно-частотных параметров речевого сигнала, тогда как фазовые характеристики мало исследованы. Это связано с двумя факторами. Во-первых, существовало мнение, что фазовые характеристики не играют существенной роли в восприятии речи. Однако постепенно было установлено, что по фазовым параметрам можно восстановить речевой сигнал [5], а фазы существенно влияют на восприятие речи [6–9]. Роль фазовых характеристик в обработке речи описывается в [10] и обзоре [11]. Второй фактор, препятствовавший использованию фазовых характеристик, заключается в том, что фазово-частотная характеристика (ФЧХ) речевого сигнала представляет собой разрывную функцию, с областью значения $[-\pi, \pi]$. Поэтому, в отличие от динамического амплитудного спектра, динамический фазовый спектр не позволяет визуально соотнести его признаки с параметрами речевого сигнала. Позднее выяснилось, что эти трудности анализа фаз можно частично обойти путем введения *групповой задержки* и *мгновенной частоты*. Групповая задержка была определена в работе [12] как отрицательное значение производной фазы по частоте в каждый момент времени. На этом понятии основаны методы нахождения моментов начала и конца активности голосового источника из работ [13, 14]. Мгновенная частота определяется как мнимая часть отношения аналитического сигнала к самому сигналу, что эквивалентно производной от фазы по времени [15]. Мгновенная частота также используется для определения моментов начала и конца активности голосового источника [16].

Экспериментальные исследования показали, что влияние фаз на восприятие речи сложным образом зависит от частоты основного тона, интенсивности сигнала и полосы частот [8, 17]. Оказывается, что влияние фаз тем больше, чем ниже частота основного тона, и это связано с ограниченной длительностью импульсов (0.5–2 мс) в нервных каналах слуховой системы. Существует также некая предельная частота, выше которой экспериментальная оценка фаз становится недостоверной [18].

Оценка фаз обычно осуществляется на основе вычисления кратковременного спектра речевого сигнала, и здесь необходимо подбирать форму и длительность взвешивающего окна в кратковременном преобразовании Фурье (КПФ). Считалось, что эта длительность при нахождении ФЧХ должна быть значительно больше, чем при вычислении амплитудного спектра. Например, в [5] предполагается, что она должна быть больше 1 с. Этот фактор будет обсужден ниже более детально (см. разд. 2).

Цель данной работы – проанализировать связь фазовых характеристик речевого сигнала с параметрами T_{op}, T_{cl} и F_0 голосового источника и предложить практические алгоритмы нахождения этих параметров из сегментов реальной речи.

Известные нам приложения фазового анализа к задачам речевых технологий носят, в значительной степени, формальный характер. В частности, мгновенная частота и групповая задержка не специфичны для анализа речи, хотя они и являются универсальными характеристиками любых сигналов. Именно поэтому в нашей статье предложены математические модели, позволяющие связать параметры голосового источника с параметрами фазово-частотных характеристик сегментов речевого сигнала (см. разд. 2). Эта связь оказывается достаточно сложной даже для простейших форм источника голосового возбуждения, и ее полный математический анализ затруднителен. Тем не менее, применение в этих моделях асимптотических методов дает возможность связать характеристики ФЧХ речевых сегментов с такими параметрами речевого источника, как его длительность, моменты начала и конца его работы (открытия и закрытия голосовой щели). Этот асимптотический анализ (разд. 2, Приложения 1 и 2) и численный анализ связи ФЧХ с параметрами источника (разд. 4) проведен в нашей работе для значительного числа синтетических и реальных речевых фрагментов, описанных в разд. 3. В результате выработаны алгоритмы, позволяющие вычислить параметры источника путем анализа ФЧХ, и эти параметры могут быть использованы в различных речевых приложениях. Рекомендации по применению алгоритмов и пределы их применимости обсуждаются в разд. 5.

2. МАТЕМАТИЧЕСКИЕ МОДЕЛИ, СВЯЗЫВАЮЩИЕ ФЧХ РЕЧЕВОГО СИГНАЛА И ГОЛОСОВОЙ ИСТОЧНИК

Рассмотрим простейшие математические модели, связывающие характеристики голосового источника с речевым сегментом и, далее, с фазово-частотной характеристикой этого сегмента. Для этого используется модель речеобразования, предложенная в работах [19, 20] и основанная на известном уравнении Вебстера. Она связывает функцию голосового источника $q(t)$ – производную объемной скорости $v(t)$ воздушного потока в

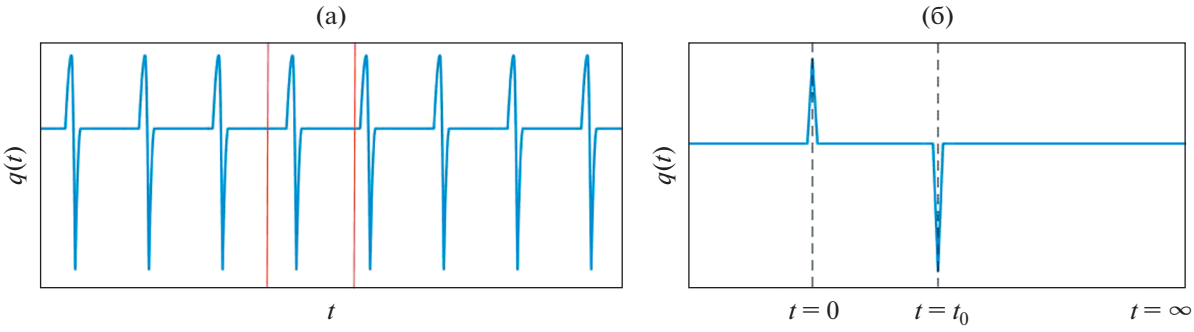

Рис. 1. (а) – Импульсы голосового источника. (б) – Выделенная вертикальными линиями часть импульсов, используемая в модели вычисления ФЧХ.

голосовой щели (ГЩ) – и генерируемый при $t \geq 0$ речевой сигнал $s(t)$ следующим образом:

$$s(t) = \int_0^t K(t-\tau)q(\tau)d\tau. \qquad (1)$$

Здесь

$$K(t) = K_0\sum_{n=1}^{N_0}\alpha_n e^{-\delta_n t}\sin\omega_n t,$$

$$\alpha_n = \left(\omega_n\frac{d\Delta}{d\omega}(\omega_n)\right)^{-1},$$

$$\Delta(\omega) = (\omega_1-\omega)\prod_{m=2}^{\infty}\frac{\omega_m-\omega}{(m-1)^2},$$

$\omega_n = 2\pi f_n$, f_n – резонансные частоты речевого тракта, δ_n – декремент затухания n-го резонанса, а K_0 – нормировочная константа, определяемая единицами измерения, которая в дальнейшем считается равной единице. Пользуясь этой моделью, можно связать некоторые числовые характеристики голосового источника с ФЧХ речевого сигнала. Это позволяет найти (оценить) по ФЧХ упомянутые характеристики.

2.1. Оценки длительности голосового источника

Примерный вид импульсов источника $q(t)$ голосового возбуждения приведен на рис. 1а. Каждый импульс (см., например, выделенный вертикальными линиями) характеризуется двумя пиками. Пик с положительной амплитудой находится вблизи момента открытия голосовой щели, а пик с отрицательной амплитудой – вблизи момента ее закрытия. Существует довольно много математических моделей, хорошо описывающих реальные голосовые источники. Однако, ни одна из них не позволяет выполнить хотя бы качественный анализ ФЧХ генерируемого сигнала. Чтобы сделать это, мы приняли идеализированную модель источника $q(t)$ в виде последовательности двух δ-образных импульсов с положительной и отрицательной амплитудами:

$$q(t) = A\delta(t) - B\delta(t-t_0), \quad A, B > 0 \qquad (2)$$

(см. рис. 1б). Здесь t_0 – время действия источника, соответствующее длине интервала открытой голосовой щели. Такая форма источника пригодна для анализа небольших по сравнению с периодом основного тона величин t_0. В этой форме неявно предполагается, что к моменту $t = 0$ произошло затухание формантных колебаний, вызванных предыдущим импульсом.

Сигнал, который генерируется источником (2) по формуле (1), имеет вид

$$s(t) = Ah(t)K(t) - Bh(t-t_0)K(t-t_0) =$$
$$= Ah(t)\sum_{n=1}^{N_0}\alpha_n e^{-\delta_n t}\sin\omega_n t - Bh(t-t_0)\times$$
$$\times\sum_{n=1}^{N_0}\alpha_n e^{-\delta_n(t-t_0)}\sin\omega_n(t-t_0),$$

где $h(t)$ – функция Хевисайда. Вычислив преобразование Фурье $\Phi(\omega) = F[s](\omega)$ этого сигнала, получим

$$\Phi(\omega) = \sum_{n=1}^{N_0}\alpha_n\left\{AF[e^{-\delta_n t}\sin\omega_n t](\omega) -\right.$$
$$\left.-BF[e^{-\delta_n t}\sin\omega_n t](\omega)e^{-i\omega t_0}\right\} = S_1(\omega)S_2(\omega).$$

Здесь

$$S_1(\omega) = A - Be^{-i\omega t_0},$$

$$S_2(\omega) = \sum_{n=1}^{N_0}\frac{\alpha_n\omega_n}{(\omega+\omega_n+i\delta_n)(\omega-\omega_n+i\delta_n)}.$$

Фазово-частотная характеристика $\varphi(\omega)$ сигнала находится из равенства

$$\ln\Phi(\omega) = \ln|\Phi(\omega)| + i\varphi(\omega) = \ln S_1(\omega) + \ln S_2(\omega) =$$
$$= \ln|S_1(\omega)| + \ln|S_2(\omega)| + i\left[\operatorname{Arg}S_1(\omega) + \operatorname{Arg}S_2(\omega)\right]_\pi$$

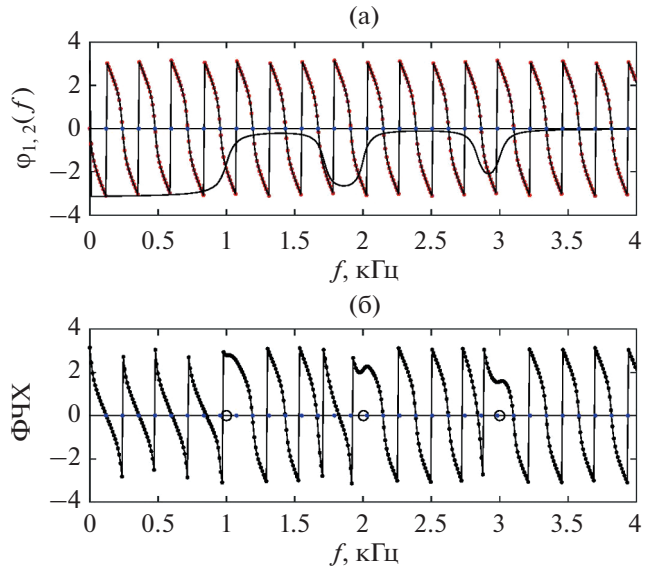

Рис. 2. (а) – Компоненты ФЧХ: периодическая часть $\varphi_1(\omega)$ – линия с точками, определяемая формантами; часть $\varphi_2(\omega)$ – непрерывная линия. (б) – ФЧХ $\varphi(\omega)$. Значения резонансных частот отмечены маркерами ○.

как

$$\varphi(\omega) = \left[\operatorname{Arg} S_1(\omega) + \operatorname{Arg} S_2(\omega)\right]_\pi = \\ = \left[\operatorname{Arg}\left(Be^{-i\omega t_0} - A\right) + \right. \\ \left. + \operatorname{Arg}\sum_{n=1}^{N_0}\frac{\alpha_n\omega_n}{(\omega+\omega_n+i\delta_n)(\omega-\omega_n+i\delta_n)}\right]_\pi. \quad (3)$$

Здесь символы $\operatorname{Arg} S_{1,2}(\omega)$ обозначают аргументы комплексных функций $S_{1,2}(\omega)$, а функция $\left[\operatorname{Arg} z\right]_\pi = \arg z$ вычисляет для комплексного числа z по величинам $\operatorname{Arg} z$ главное значение аргумента, лежащее в пределах от $-\pi$ до π. Отметим, что эта функция не обладает свойством аддитивности: вообще говоря,

$$\left[\operatorname{Arg} S_1(\omega) + \operatorname{Arg} S_2(\omega)\right]_\pi \neq \\ \neq \left[\operatorname{Arg} S_1(\omega)\right]_\pi + \left[\operatorname{Arg} S_2(\omega)\right]_\pi,$$

т.е. фазово-частотная характеристика сигнала не складывается непосредственно из фаз величин $S_1(\omega)$, $S_2(\omega)$. Это осложняет анализ равенства (3). Тем не менее, ясно, что функция $\varphi(\omega)$ имеет разрывы 1 рода в точках ω, в которых величина $\operatorname{Arg} S_1(\omega) + \operatorname{Arg} S_2(\omega)$ принимает значения, кратные $\pm\pi$.

Проведем более детальный анализ. Установим связь параметра источника t_0 со свойствами ФЧХ. Сначала формально рассмотрим случай $\delta_n = 0$ (отсутствие потерь в речевом тракте). Тогда величина

$$S_2(\omega) = \sum_{n=1}^{N_0}\frac{\alpha_n\omega_n}{(\omega+\omega_n+i\delta_n)(\omega-\omega_n+i\delta_n)} = \\ = \sum_{n=1}^{N_0}\frac{\alpha_n\omega_n}{(\omega+\omega_n)(\omega-\omega_n)}$$

действительная. Поэтому из (3) следует:

$$\varphi(\omega) = \varphi_1(\omega) = \left[\operatorname{Arg}\left(Be^{-i\omega t_0} - A\right)\right]_\pi = \\ = \arg\left(Be^{-i\omega t_0} - A\right) = \\ = \arg\left((B\cos\omega t_0 - A) - iB\sin\omega t_0\right).$$

Типичный вид функции $\varphi(\omega) = \varphi_1(\omega)$ показан на рис. 2 сверху линией с точками. Функция $\varphi(\omega)$ равна нулю или имеет точку разрыва при тех ω, для которых мнимая часть числа $Be^{-i\omega t_0} - A$ обращается в нуль, т.е. при $\sin\omega t_0 = 0$. Все такие точки имеют вид: $\omega = \Omega_m = \dfrac{\pi m}{t_0}$, $m = 0, \pm1, \pm2, \ldots$, и соответствующие частоты выражаются как $f_m = \dfrac{m}{2t_0}$. Расстояния между этими характерными точками, т.е. числа $\Delta f_m = f_{m+1} - f_m = \dfrac{1}{2t_0}$, определяют период ФЧХ. Поэтому в идеализированном варианте, найдя нули и точки разрыва ФЧХ, мы можем вычислить параметр источника как $t_0 = \dfrac{1}{2\Delta f_m}$.

В случае малых потерь в речевом тракте можно провести похожий анализ. Он дан в Приложении 1. В этом случае величины $\Delta f_m = f_{m+1} - f_m$ будут уже зависеть от m. Однако, формула для t_0 остается верной в виде $t_0 = \dfrac{1}{2\overline{\Delta f_m}}$, где $\overline{\Delta f_m} = \lim_{N\to\infty}\dfrac{1}{N}\sum_{m=1}^{N}\Delta f_m$ – среднее значение величин Δf_m. Таким образом, вычислив приближенно это среднее значение из ФЧХ для частот, больших первой форманты, мы можем оценить параметр источника t_0 по той же формуле, что и в случае отсутствия потерь, в котором $\overline{\Delta f_m} = \Delta f_m$.

На рис. 2 показан пример ФЧХ, а также связанные с ней функции $\varphi_1(\omega) = \left[\operatorname{Arg} S_1(\omega)\right]_\pi$, $\varphi_2(\omega) = \left[\operatorname{Arg} S_2(\omega)\right]_\pi$ (рис. 2а). ФЧХ (рис. 2б) вычислена по формуле (3) с $A = 1$, $B = 1.5$ для резонансных частот $F_{1-3} = [1, 2, 3]$ кГц и величин $\delta_n = 2\pi[0.05, 0.04, 0.05]$, $\alpha_n = [1, 0.5, 0.25]$ при $t_0 = 4.18$ мс. Численно найдя величину $\overline{\Delta f_m}$ для этой ФЧХ в частотном диапазоне $f \in (1, 4)$ кГц, получим оценку параметра источника: $t_0 \approx 4.17$ мс.

Численные эксперименты по верификации равенства $t_0 = 0.5\left(\overline{\Delta f_m}\right)^{-1}$ для реальных речевых сигналов представлены ниже в разд. 4.1.

2.2. Вычисление ФЧХ для конечного периода основного тона

Откажемся от сделанного в разд. 2.1 предположения о том, что период основного тона T_0 много больше длительности t_0 открытой ГЩ. Тогда для источника с M одинаковыми периодами основного тона можно принять модель

$$q(t) = \sum_{m=0}^{M-1}\left[A\delta(t - mT_0) - B\delta(t - t_0 - mT_0)\right],$$
$$A, B > 0.$$

Согласно (1), модельный сигнал приобретет вид

$$s(t) = \int_0^t K(t-\tau)q(\tau)d\tau =$$
$$= \sum_{m=0}^{M-1}\left[Ah(t - mT_0)K(t - mT_0) - \right.$$
$$\left. -Bh(t - t_0 - mT_0)K(t - t_0 - mT_0)\right],$$

и его преобразование Фурье вычисляется так:

$$\Phi(\omega) = F[s](\omega) = \tilde{K}(\omega)\left[A - Be^{-i\omega t_0}\right]\frac{1 - e^{-i\omega MT_0}}{1 - e^{-i\omega T_0}},$$

$$\tilde{K}(\omega) = \sum_{n=1}^{N_0}\frac{\alpha_n\omega_n}{(\omega + \omega_n + i\delta_n)(\omega - \omega_n + i\delta_n)}$$

(см. Приложение 2). Для простоты рассмотрим случай двух периодов основного тона источника ($M = 2$) при условии $\delta_n = \delta \ll \omega_1$ малых потерь в тракте, которое обычно выполнено, и при "больших частотах" ($\delta \ll \omega$). Тогда (см. Приложение 2)

$$\Phi(\omega) = \left[A - Be^{-i\omega t_0}\right]\left(1 + e^{-i\omega T_0}\right)\sum_{n=1}^{N_0}\frac{\alpha_n\omega_n}{(\omega^2 - \omega_n^2)}.$$

Можно видеть, что даже при таких упрощениях аналитическое исследование поведения ФЧХ сигнала, т.е. функции

$$\varphi(\omega) = \mathrm{Im}\left\{\ln\Phi(\omega)\right\} =$$
$$= \mathrm{Im}\left\{\ln(A - Be^{-i\omega t_0}) + \ln(1 + e^{-i\omega T_0})\right\} \quad (4)$$

в зависимости от ω и параметров t_0, T_0 затруднительно. Поэтому приходится проводить исследование численно. Приведем пример такой ФЧХ, вычисленной для $A = 1$, $B = 1.5$, формантных частот $F = [1, 2, 3]$ кГц и параметров $t_0 = 2$ мс, $T_0 = 10$ мс, $\alpha_n = [1,\ 0.5, 0.25]$, $\delta = 0.15$.

Из рис. 3 и формулы (4) видно, что фазовая функция определяется двумя колебаниями. Одно

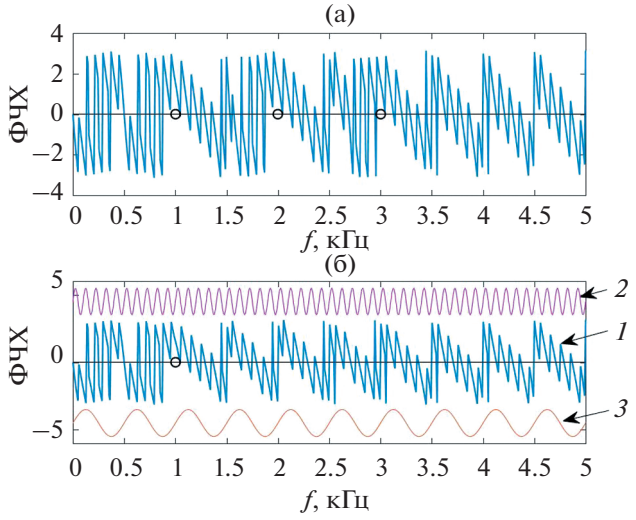

Рис. 3. (а) – ФЧХ для трех формант, положение которых отмечено кружками. (б) – *1* – ФЧХ для одной форманты с частотой 1 кГц; *2* – гармоника, определяемая частотой основного тона $F_0 = 1/T_0$; *3* – гармоника, определяемая частотой источника $f_0 = 1/t_0$.

с периодом t_0 связано с длительностью источника возбуждения, другое происходит с периодом основного тона T_0. Последнее свойство будет использоваться ниже (см. разд. 4) в экспериментах по определению частоты основного тона синтетических и реальных речевых сигналов. В Приложении 2 рассмотрен вопрос о нулях ФЧХ в рассматриваемом случае конечного периода основного тона. Показано, что одна из возможных серий нулей имеет вид $\omega_k = 2\pi f_k = \pi(1 + 2k)T_0^{-1}$, и она порождает серию величин $\Delta f_k = f_{k+1} - f_k = T_0^{-1}$.

2.3. Оценка параметров голосового источника

При обработке реальных дискретных речевых сигналов вместо стандартного преобразования Фурье часто используется кратковременное преобразование Фурье (КПФ)

$$\Phi(f, t_c) = \int_0^{+\infty} e^{-i2\pi ft}w(t - t_c)s(t)dt. \quad (5)$$

Здесь $w(t - t_c)$ – задаваемое пользователем окно преобразования с центром t_c. В этом случае для нахождения ФЧХ используется формула $\varphi(f, t_c) = \mathrm{Im}\left\{\ln\Phi(f, t_c)\right\}$, и вместо одной фазовой функции получается их семейство, зависящее от t_c. Оказывается, что, анализируя изменения этих ФЧХ в зависимости от положения центра окна по отношению к сигналу, можно оценить моменты

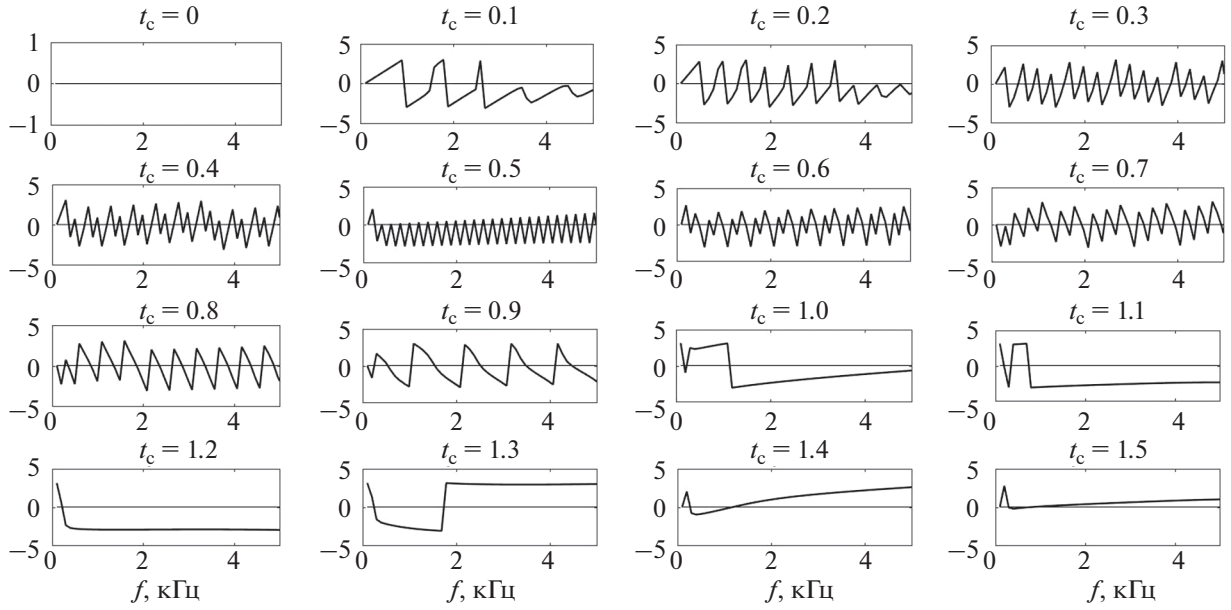

Рис. 4. ФЧХ $\varphi(f, t_c)$ сигнала в окне с центром t_c.

включения и выключения голосового источника. Рассмотрим эти изменения на примере сигнала

$$s(t) = Ah(t - T_{op})e^{-\delta(t-T_{op})}\sin 2\pi f(t - T_{op}),\ [t] = \text{мс},$$

соответствующего открытию ГЩ в момент T_{op}. Для того чтобы можно было провести аналитические вычисления, упрощающие выражение (5), выберем модельное окно с полушириной 0.5 мс в форме $w(t) = \{\sin(\pi t),\ 0 \le t \le 1;\ 0, (t < 0) \cup (t > 1)\}$. На рис. 4 приведены фазовые функции $\varphi(f, t_c)$, вычисленные для такого сигнала при $f = 1.2$ кГц, $A = 1$, $\delta = 0.04\pi$ и $T_{op} = 0.5$ мс для различных времен $t_c, 0 \le t_c \le 1.5$ мс.

Сопоставим ФЧХ, показанные на рис. 4, с соответствующими положениями сигнала в окне, т.е. с видом функций времени $\zeta(t, t_c) = w(t - t_c)s(t)$ для различных t_c (см. рис. 5).

Модельный сигнал отличен от нуля при $t > T_{op} = 0.5$ мс. Поэтому первоначально (при $t_c = 0$) момент T_{op} не попадает в окно полуширины 0.5 мс, и поэтому ФЧХ обращается в нуль. При $t_c > 0$ момент T_{op} входит в окно, и ФЧХ приобретает колебательный характер ($t_c = 0.1$–0.9 мс). Частота ее колебаний увеличивается, пока центр окна t_c не совпадет с $T_{op}(t_c = T_{op} = 0.5$ мс). Затем эта частота уменьшается, пока при $t_c > 1$ мс точка T_{op} не выходит из окна, и колебания ФЧХ практически пропадают. Квазипериодический характер по f каждой из этих ФЧХ можно охарактеризовать частотой появления их нулей f_m, т.е. величиной

$$Q_1(t_c) = \overline{\Delta f_m}(t_c) = \frac{1}{N}\sum_{m=1}^{N}\Delta f_m.$$

В дальнейшем мы будем называть функцию $Q_1(t_c)$ кривой квазипериодов первого типа для ФЧХ. Можно также охарактеризовать квазипериодичность фазово-частотной характеристики $\varphi(f, t_c)$ с помощью функции $Q_2(t_c) = \max_m\{\Delta f_m\}$ – кривой квазипериодов второго типа. Кривые квазипериодов $Q_1(t_c)$, рассчитанные для сигналов $s(t) = Ah(t - T_{op})e^{-\delta(t-T_{op})}\sin 2\pi f(t - T_{op})$ с различными частотами: $f = [1.2, 2.2, 3.2]$ кГц, показаны на рис. 6а. На рис. 6б изображены кривые $Q_2(t_c)$, вычисленные аналогичным образом.

Видно, что обе кривые, $Q_1(t_c)$ и $Q_2(t_c)$, имеют выраженный локальный минимум при $t_c = T_{op}$, т.е. при совпадении центра окна КПФ с моментом открытия ГЩ. Аналогичный вид имеют зависимости $Q_1(t_c)$, $Q_2(t_c)$ вблизи момента T_{cl} закрытия ГЩ.

Приведенные примеры не учитывают дискретизацию сигнала, характерную для реально регистрируемой речи. На рис. 7 показано, как выглядят типичные кривые квазипериодов при дискретизации с частотой 16 кГц.

Влияние дискретизации выражается в появлении высокочастотных осцилляций кривых. После фильтрации этих осцилляций кривые квази-

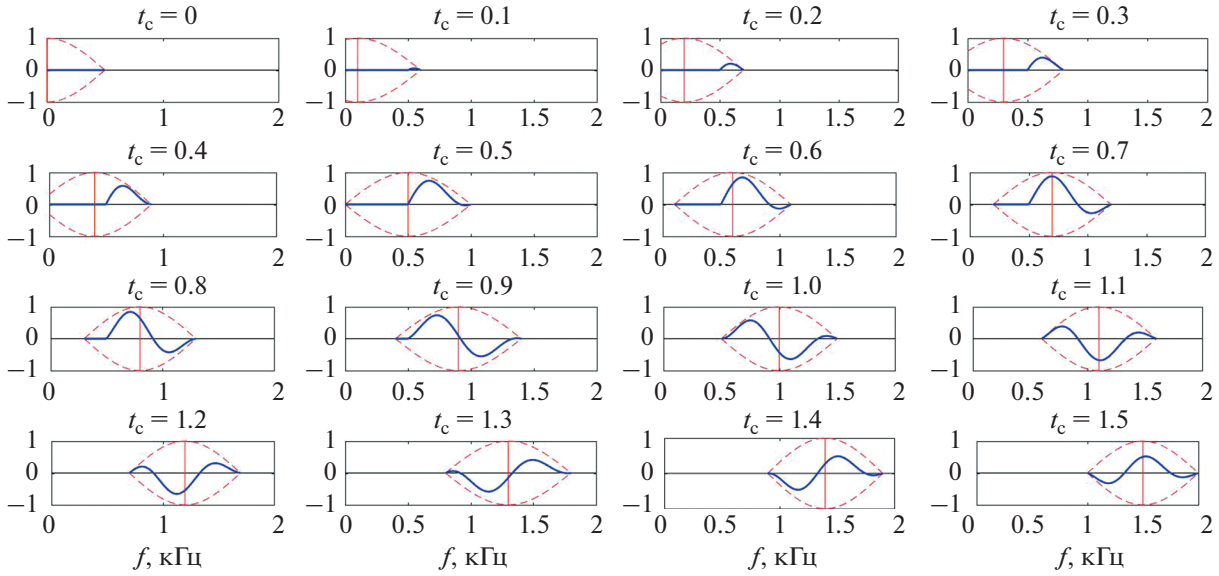

Рис. 5. Сигнал в окне с центром t_c: $\zeta(t,t_c) = w(t - t_c)s(t)$. Непрерывная линия – функция $\zeta(t,t_c)$; пунктир – окно; вертикальная линия – центр окна t_c.

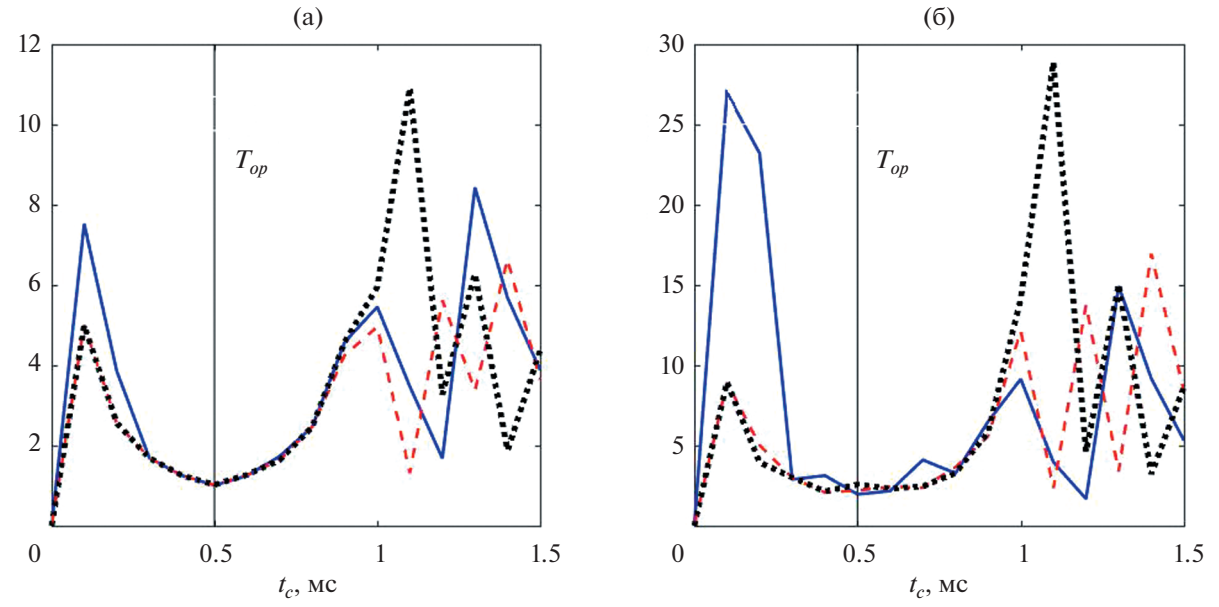

Рис. 6. (а) – Кривые квазипериодов $Q_1(t_c)$ при положении окна КПФ вблизи момента открытия ГЩ; непрерывная линия – сигнал с f = 1.2кГц, пунктир – с f = 2.2 кГц, точки – с f = 3.2кГц. (б) – Аналогичные кривые квазипериодов $Q_2(t_c)$.

периодов приобретают формы, схожие с представленными на рис. 6.

Вид кривых квазипериодов наводит на мысль о том, что их минимумы можно использовать для определения моментов открытия и закрытия ГЩ при разметке реального речевого сигнала.

Сделаем некоторые выводы из рассмотрения предложенных моделей ФЧХ.

1. Существуют две колебательные компоненты ФЧХ, период одной из которых связан с длительностью источника возбуждения, а период другой – с периодом основного тона (рис. 3).

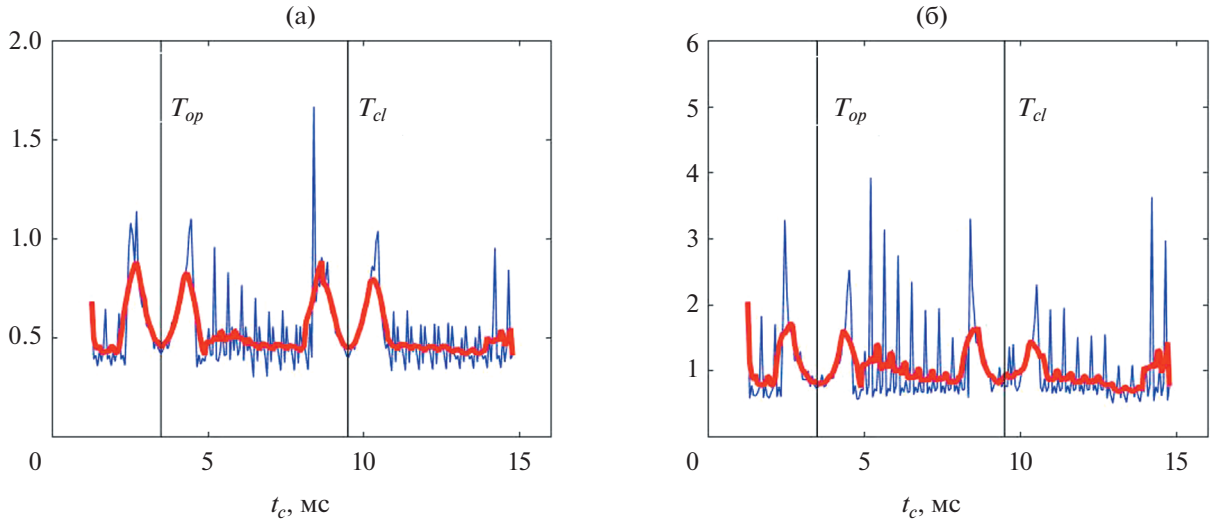

Рис. 7. Кривые квазипериодов (тонкая линия) для дискретного сигнала вблизи моментов открытия и закрытия ГЩ (вертикальные линии): (а) – $Q_1(t_c)$; (б) – $Q_2(t_c)$. Жирной линией показаны те же кривые после фильтрации (усреднения с помощью скользящего среднего).

2. Частоты, на которых наблюдается нарушение периодичности фазово-частотной функции, связаны с резонансными частотами мод (рис. 2б). Нарушение периодичности может возникнуть и для тех ФЧХ, у которых сигнал в окне КПФ не содержит момент включения источника (рис. 4, 5).

3. Моменты начала и конца активности источника возбуждения находятся вблизи минимумов кривых квазипериодов (рис. 6, 7).

На основе этих выводов было выполнено исследование свойств ФЧХ синтетических и реальных речевых сигналов. Реальный голосовой источник существенно отличается от использованных нами в модели. Часто это выражалось в том, что для сегментов речи минимумы кривых квазипериодов оказывались плохо обусловленными, и соответствующие оценки моментов T_{cl} и T_{op} становились ненадежными. Поэтому методы оценки параметров голосового источника пришлось несколько скорректировать в процессе численных экспериментов. В частности, наименьшая погрешность оценки искомых параметров была получена для эвристического алгоритма, состоящего в поиске экстремумов функции

$$\theta(t_c) = \frac{\Delta f_{\max}}{M(\Delta f)} \qquad (6)$$

вместо функций квазипериодов. Здесь $\Delta f_{\max}$ – максимальный интервал между нулями фазовой функции $\varphi(f, t_c)$, а

$$M(\Delta f) = \frac{1}{N-1}\left(\sum_{m=1}^{N} \Delta f_m - \Delta f_{\max}\right).$$

В численных экспериментах было найдено, что функция $\theta(t_c)$ имеет минимум вблизи момента T_{cl}, как и для описанных выше кривых квазипериода. Однако, вблизи момента T_{op} эта функция имеет максимум. Результаты соответствующих численных экспериментов по определению частоты основного тона и моментов начала и конца активности голосового источника приведены в разд. 4.2 и 4.3. В них функция $\theta(t_c)$ обозначается как $\theta(t)$.

3. БАЗЫ РЕЧЕВЫХ ДАННЫХ

Оценка эффективности метода определения параметров голосового источника требует знания истинных значений этих параметров. Наиболее распространенный подход для такой оценки состоит в анализе сигналов, синтезированных с заданным голосовым источником. Другой подход использует косвенные оценки параметров голосового источника путем измерения каких-либо физических характеристик, связанных с активностью голосового источника на реальных речевых сегментах. Здесь, в частности, можно использовать так называемые глоттограммы, т.е. измерения напряжения между поверхностными электродами, наложенными симметрично по обе стороны щитовидного хряща. В наших численных экспериментах использовались данные обоих типов. Им соответствовали базы данных, содержащие сигналы трех видов.

База 1. Сигналы, синтезированные по параметрам 6 русских гласных /*а*, *э*, *и*, *ы*, *о*, *у*/ в диапазоне частот основного тона от 80 до 380 Гц. Один из параметров, характеризующих импульс источ-

Таблица 1. Параметры гласных звуков, Гц

а	F	600	1200	2300	3500	3806	4742
	ΔF	80	50	80	100	150	220
о	F	500	910	2320	2630	4030	4730
	ΔF	100	50	70	90	140	190
у	F	408	860	2040	2760	3610	4430
	ΔF	150	40	50	70	90	120
и	F	290	2272	3100	4000	5050	6110
	ΔF	150	40	50	70	90	120
ы	F	286	1874	2570	3730	4420	5050
	ΔF	150	42	54	71	92	120
э	F	490	1350	2230	2770	3670	4230
	ΔF	70	40	60	80	110	140

ника голосового возбуждения, определяет отношение длительности импульса к периоду основного тона T_0, $OQ = (T_{cl} - T_{op})/T_0$. Ранее в [2] на материале базы Arctic [21] было установлено, что распределение величин OQ находится в диапазоне 0.25–0.8. С целью проверки влияния этого фактора на ошибки оценок моментов начала и конца импульса источника синтезировались сигналы с OQ от 0.2 до 0.8 с шагом 0.2. Формантные частоты F и ширина полосы каждого резонанса ΔF представлены в Табл. 1.

Для синтеза речевых сигналов использовался источник возбуждения с пятью параметрами, описанный в [20]:

$$q(t) = \begin{cases} \sin\dfrac{\pi t}{2T_1}, \ 0 \le t \le T_1; \ \ (A_0 + 1)\cos\dfrac{\pi(t - T_1)}{2(T_2 - T_1)} - A_0, \ T_1 \le t \le T_2; \\ -A_0\dfrac{(T_3 - t)^{2\gamma}}{(T_3 - T_2)^{2\gamma}}, \ T_2 \le t \le T_3; \ 0, \ T_{cl} \le t \le T_0. \end{cases} \quad (7)$$

Здесь T_0 – период основного тона, T_1 и T_2 – моменты максимального и минимального значения источника возбуждения, $T_3 = T_{cl}$ – момент окончания действия источника, параметр γ определяет скорость закрытия голосовой щели. Величина A_0 определяется из равенства нулю объемной скорости воздушного потока через голосовую щель в момент ее закрытия: $\int_{T_{op}}^{T_{cl}} q(t)dt = 0$ и вычисляется как $A_0 = 2T_2((T_2 - T_1)(\pi - 2) + \pi(2\gamma + 1)(T_3 - T_2))^{-1}$. Для этой модели $OQ = (T_3 - T_{op})/T_0$.

Отметим, что в базе 1 точно известны моменты начала и конца действия источника возбуждения. Речевой сигнал синтезировался фильтром, сконструированным по методу линейного предсказания.

База 2. Сигналы из Repository3, представленные по ссылке из статьи [22]. Акустические сигналы были получены с использованием трехмерной физической модели речевого тракта для мужского и женского голосов и гласных /*a*, *e*, *i*, *u*/. Эта модель возбуждалась параметрическим голосовым источником *LF* с четырьмя параметрами [23] с частотами основного тона от 100 до 380 Гц с $OQ \approx 0.36$. Наряду с синтезированными сигналами, в Repository3 имеются и сигналы, соответствующие импульсам источника возбуждения. В экспериментах с этими сигналами моменты начала и конца импульса определялись численно как моменты его обращения в нуль.

База 3. В этой базе представлены сигналы из базы Arctic [21]. Имеются записи голосов трех дикторов – двух мужчин, обозначенных как BDL и JMK, и одной женщины (SLT), произносивших около 1100 фраз. Записи сигналов производились в заглушенной камере с одновременной регистрацией глоттограмм. В экспериментах с этой базой моменты начала и конца импульса источника возбуждения определялись соответственно по максимумам и минимумам производной глоттограммы. Эти параметры также использовались для оценки частоты основного тона.

Сигналы в базе 1 русских гласных синтезировались с частотой отсчетов 16 кГц, а сигналы из баз 2 и 3 были пересчитаны на эту частоту. В экспери-

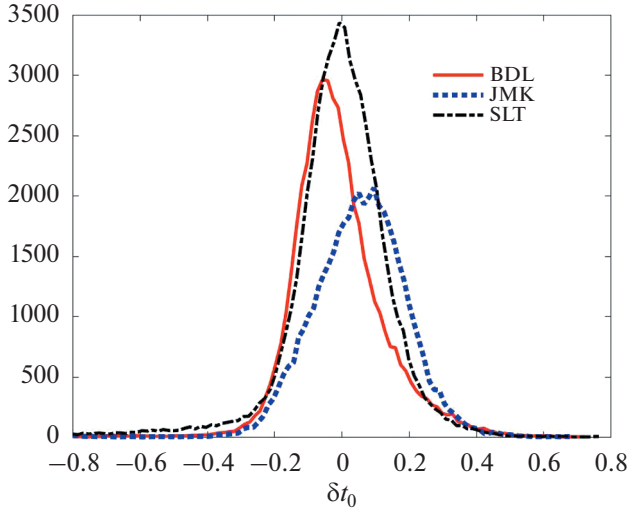

Рис. 8. Эмпирические распределения относительных ошибок $\delta t_0 = \left(t_0 - t_{0\,\mathrm{exp}}\right)\big/t_{0\,\mathrm{exp}}$ оценки величин $t_{0\,\mathrm{exp}}$ реальных длительностей работы ГИ для трех дикторов из базы 3.

ментах использовалось кратковременное преобразование Фурье с окном Гаусса $w(t) = \exp(-t^2/a^2)$ с параметром $a = 2.5$ при анализе частоты основного тона, и $a = 1$ при оценке моментов начала и конца импульса источника возбуждения. При оценке периода основного тона длительность окна составляла 16 мс, а при оценке моментов T_{op} и T_{cl} – 2.5 мс. Эти величины были найдены экспериментально. Они существенно расходятся с общепринятыми рекомендациями о необходимости использования большой длительности окна, как это упоминается во Введении.

4. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

4.1. Оценка длительности открытой голосовой щели

В этой серии вычислительных экспериментов изучалось, насколько точно реальная длительность открытой голосовой щели может быть оценена по полученной в разд. 2.1 простой формуле $t_0 = 0.5\left(\overline{\Delta f_m}\right)^{-1}$. В экспериментах использовалась база 3, содержащая реальные речевые сигналы. В каждом речевом сегменте этой базы начало и конец импульсов голосового источника определялись по параметрам глоттограмм. Поэтому можно сравнить экспериментальные длительности открытой ГЩ $t_{0\,\mathrm{exp}}$ и теоретические величины t_0, найденные из ФЧХ для каждого периода основного тона. Именно это было сделано для всех речевых сегментов трех дикторов из базы 3. Полученные эмпирические распределения относительных ошибок $\delta t_0 = \left(t_0 - t_{0\,\mathrm{exp}}\right)\big/t_{0\,\mathrm{exp}}$ оценки t_0 для экспериментальных величин $t_{0\,\mathrm{exp}}$ представлены на рис. 8.

Таблица 2. Характеристики распределений ошибки δt_0 определения параметра t_0

Диктор	N_{fp}	$\overline{\delta t_0}$	$\sigma(\delta t_0)$
BDL	48706	−0.25	13.4
JMK	43828	6.22	13.5
STL	57139	0.27	14.6

Вычисленные распределения характеризуются величинами, приведенными в табл. 2. В ней N_{fp} – количество периодов основного тона, использованных для фазового анализа, $\overline{\delta t_0}$ – средние значения ошибки δt_0, $\sigma(\delta t_0)$ – среднеквадратичные уклонения ошибки.

Из таблицы видно, что в среднем теоретическая величина t_0 удовлетворительно описывает экспериментальные данные $t_{0\,\mathrm{exp}}$: средняя ошибка – около 1–7% со среднеквадратичным уклонением около 14%.

В расчетах величин t_0 с использованием ФЧХ (см. разд. 2.1) применялось дискретное преобразование Фурье на каждом периоде основного тона. Поэтому такой подход в вычислении t_0 по сигналу с частотой отсчетов 16 кГц пригоден в основном для периодов сравнительно большой длительности (5–10 мс), т.е. для относительно малых частот основного тона (0.1–0.2 кГц). Такое условие, однако, выполнено для значительной части речевых сегментов дикторов из базы Arctic. Убедиться в этом можно по графикам распределения частот основного тона для каждого диктора (см. ниже рис. 10). По этой причине ошибки нахождения величин t_0, приведенные в табл. 2, и оказались относительно малыми.

4.2. Оценка частоты основного тона

Оценка периода основного тона определяется по введенной выше функции $\theta(t)$ как среднее расстояние между точками локальных максимумов функции $\theta(t)$ на всем сегменте гласного. На рис. 9 представлены средние по всем гласным относительные ошибки оценки частоты основного тона для различных речевых сигналов: из базы 1 с источником вида (7), из базы 2 с *LF*-источником и для сигналов из базы 3. Выяснилось, что ошибки в экспериментах с базой 1 мало зависят от отношения *OQ*, и поэтому было выполнено усреднение по всем *OQ*.

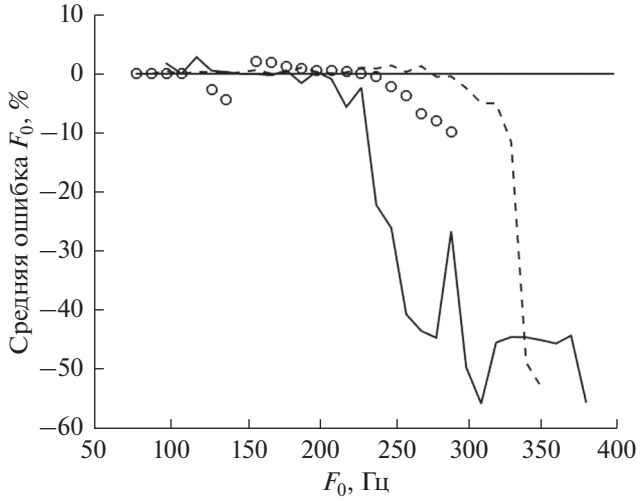

Рис. 9. Ошибки оценки частоты основного тона синтезированных русских гласных (---), сигналов из базы 2 (–) и сигналов из базы 3 (ooo).

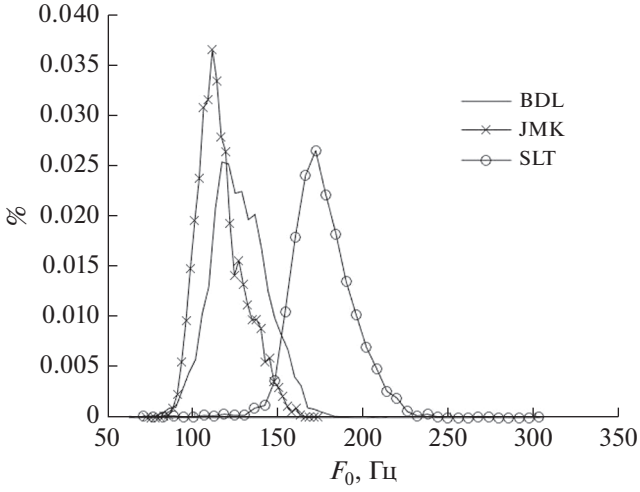

Рис. 10. Распределение частот основного тона дикторов из базы 3.

Видно, что существует некоторая критическая частота основного тона, выше которой ошибки резко возрастают. Если положить порог ошибки равным −5%, то эта частота для данных базы 2 близка к 230 Гц, несколько выше (около 260 Гц) для базы 3, а для русских гласных из базы 1 – около 320 Гц. В диапазоне частот ниже 200 Гц ошибка оказалась порядка 0.1%. Это значительно ниже по сравнению с наиболее успешным, по нашему мнению, алгоритмом [24], где ошибка достигала ±8% на этих же данных.

Ограниченная представительность частот основного тона в базе 3 определяется распределением оценок по параметрам глоттограмм (рис. 10). Наиболее вероятные значения частоты основного тона мужских голосов (дикторы BDL и JMK) оказались близки к 116 и 111 Гц, а у женского голоса (диктор SLT) к 171 Гц.

4.3. Оценки моментов начала и конца импульса голосового источника

Выше отмечалось, что при поиске моментов начала и конца импульса голосового источника T_{op} и T_{cl} весьма важна “настройка” параметров КПФ. В нашем случае, наиболее подходящим для спектрального анализа оказалось окно Гаусса длиной 2.5 мс с параметром $a = 1$. В экспериментах было установлено, что максимумы функции $\theta(t)$ находятся вблизи моментов T_{op}, а ее минимумы – вблизи величин T_{cl}, как для синтезированных сигналов, так и для сигналов, сгенерированных физической моделью речевого тракта. Для иллюстрации представим рис. 11. На нем вверху показана последовательность импульсов “объемной скорости” голосового источника, следующих с частотой $F_0 = 100$ Гц на интервале времени в 0.1 с для данных из базы 2. Внизу дан для сравнения график соответствующей функции $\theta(t)$.

Расхождение между оценками и истинными значениями моментов T_{op} и T_{cl} будем характеризовать средней ошибкой по отношению к периоду основного тона T_0 на всей длительности сегмента речи. Средняя ошибка для базы 2 по всем гласным для мужского и женского голосов в зависимости от частоты основного тона оказалась удовлетворительной: относительная погреш-

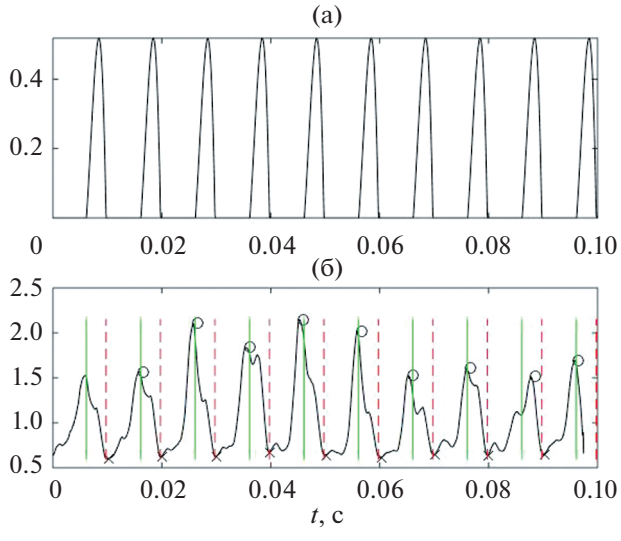

Рис. 11. (а) – Нормированная объемная скорость голосового источника. (б) – Функция $\theta(t)$. Оценка моментов начала (o) и конца источника (x). Сплошная вертикальная линия обозначает истинные моменты начала источника, а пунктирная – конца источника.

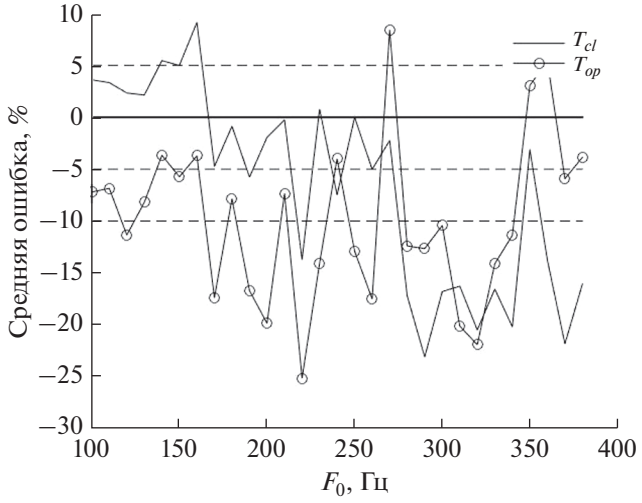

Рис. 12. Средние относительные ошибки для данных базы 2. Пунктиром показаны значения ошибок ±5 и ±10%.

ность в большинстве случаев не превышает 5%, и всегда ниже 10% (рис. 12). Однако разброс ошибок по импульсам внутри сегмента гласного может быть весьма велик, достигая 20% и более для некоторых значений частоты основного тона.

Сигналы в базе 2 были сгенерированы практически с фиксированным значением параметра *OQ*, тогда как прямые измерения воздушного потока через голосовую щель указывают на определенное разнообразие этого параметра. В отличие от оценок частоты основного тона, эксперименты по определению величин T_{op} и T_{cl} с сигналами из базы 1 выявили сильную зависимость их оценок и от F_0, и от параметра *OQ*. Зависимости средних относительных ошибок этих оценок от F_0 приведены на рис. 13а, 13б для различных *OQ*.

Сигналы в базе 3 позволяют оценить разброс ошибок определения моментов начала и конца работы голосового источника для разных дикторов. На рис. 14 показаны распределения этих ошибок, усредненные по всем произнесениям для каждого диктора из базы 3. На этих распределениях видно, что существует заметная доля ошибок с положительным или отрицательным знаком относительно наиболее вероятного значения.

Распределения на графиках не являются унимодальными. Для таких распределений наиболее вероятная ошибка более адекватно оценивает свойства распределения по сравнению со средней ошибкой, которая может оказаться близкой к нулю. Это наблюдалось и для распределений ошибок по всем частотам основного тона, где всплески положительных и отрицательных ошибок при оценке среднего значения компенсируют друг друга.

В табл. 3 представлены наиболее вероятные относительные ошибки $\delta_{\max}(T_{op})$ и $\delta_{\max}(T_{cl})$ оценок моментов открытия и закрытия голосовой щели вместе со среднеквадратическими отклонениями $\sigma(T_{op})$ и $\sigma(T_{cl})$ этих оценок. В численных экспериментах было обнаружено, что наиболее вероятная ошибка зависит от частоты основного тона, причем существует критическая частота основного тона, примерно равная 220 Гц, выше которой ошибка быстро возрастает. Отметим также, что зависимости ошибок от частоты основного тона отличаются для разных дикторов (см. рис. 15а, 15б).

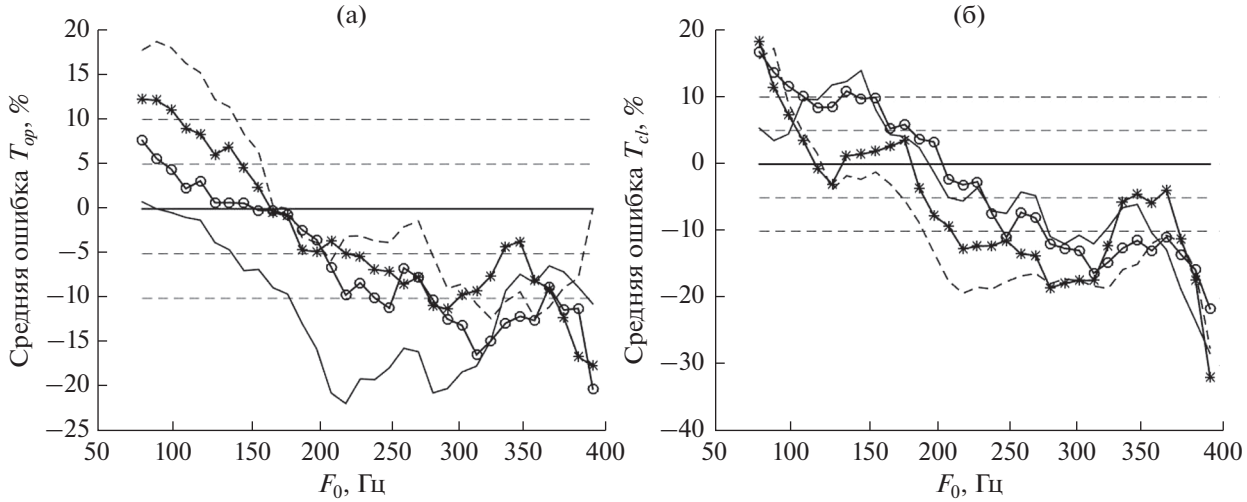

Рис. 13. Средние относительные ошибки оценки моментов (а) – T_{op} и (б) – T_{cl}. Значения параметра *OQ* – 0.2 (–); 0.4 (–o–); 0.6 (–*–); 0.8 (---). Пунктиром размечены значения ошибок ±5 и ±10%.

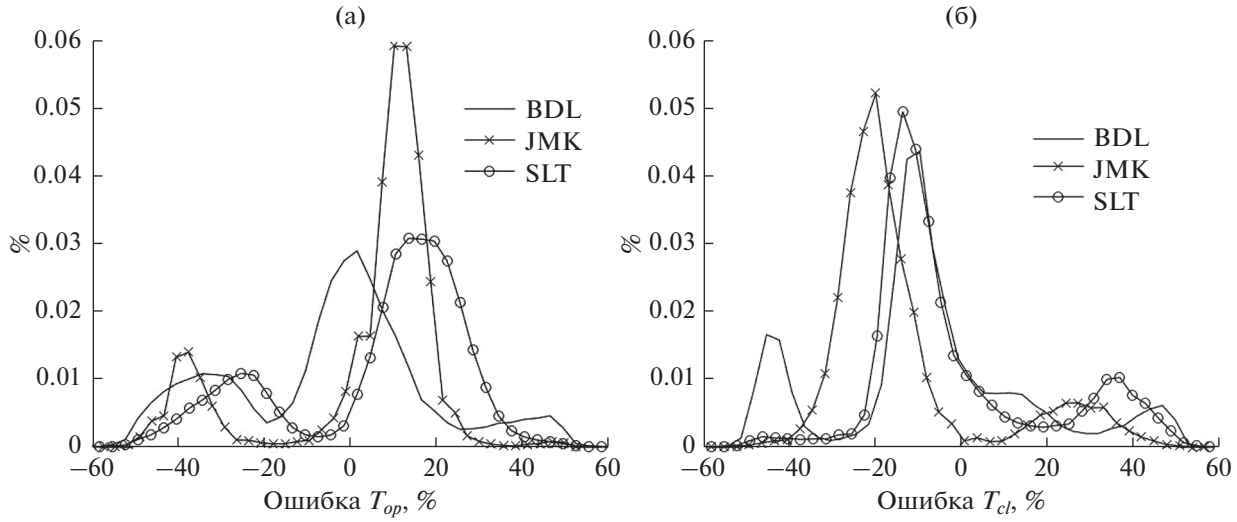

Рис. 14. Распределение средних ошибок оценок моментов (а) – T_{op} и (б) – T_{cl}.

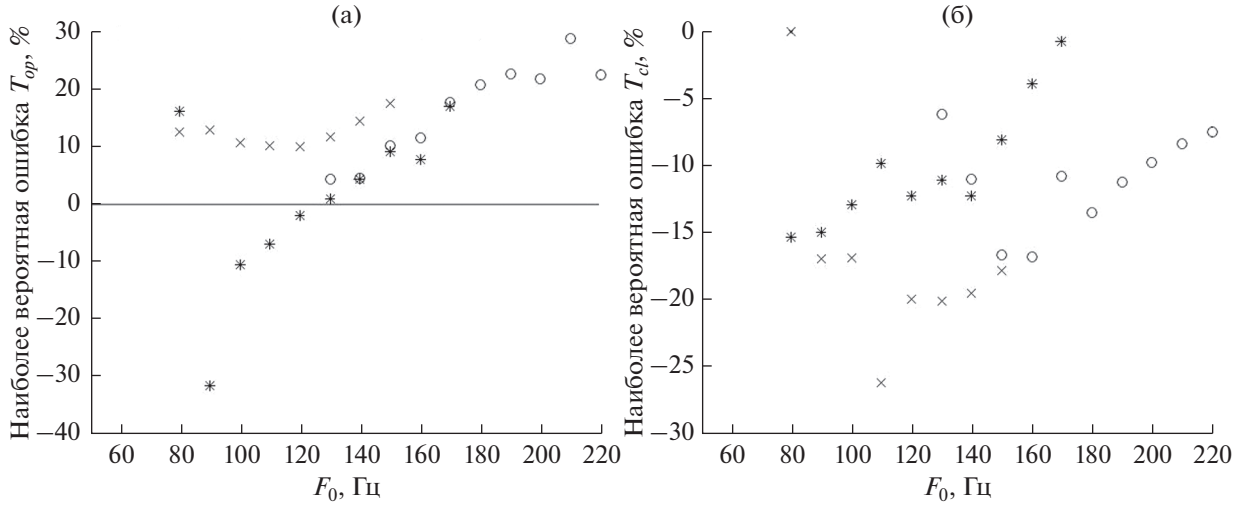

Рис. 15. Ошибки оценок моментов (а) – T_{op} и (б) – T_{cl}. Диктор BDL представлен маркерами (x), диктор JMK – маркерами (*), а диктор SLT – маркерами (○).

Моменты открытия и закрытия голосовой щели могут быть определены не только по максимальным и минимальным значениям функции $\theta(t)$. Информация об этих моментах также содержится и в значениях частоты $\phi(t)$, с которой начинается наиболее длительный интервал между нулями дискретной фазовой функции. Максимальное и минимальное значение этой частоты сложным образом зависит от резонансных частот речевого тракта.

На рис. 16 представлен речевой сигнал и различные функции, используемые при оценке моментов открытия и закрытия голосовой щели для пятой гласной в первой фразе "*Author of the danger trail ...*", произнесенной диктором BDL из базы 3. Здесь средняя ошибка определения момента T_{op} по функции $\theta(t)$ составляет –7.4, и –8.2% для момента T_{cl}. На рис. 16б показана функция, экстремумы которой используются для определения периода основного тона.

Таблица 3. Наиболее вероятная ошибка определения моментов открытия и закрытия голосовой щели, %. База 3

Диктор	$\delta_{max}(T_{op})$	$\delta_{max}(T_{cl})$	$\sigma(T_{op})$	$\sigma(T_{cl})$
BDL	1.5	–9.9	0.02	0.04
JMK	10.2	–20.2	0.04	0.04
SLT	13.5	–13.9	0.03	0.03

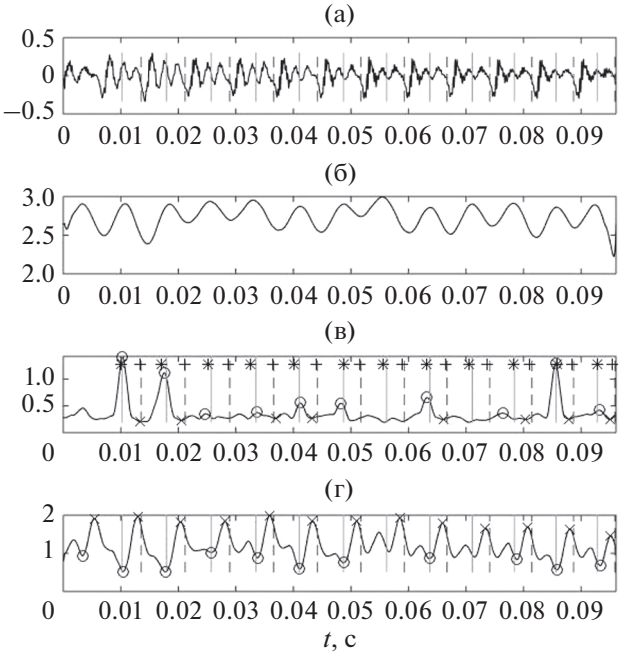

Рис. 16. (а) – Осциллограмма звукового давления, (б) – функция $\theta(t)$ с окном 16 мс, (в) – функция $\theta(t)$ с окном, равным $0.5T_0$, (г) – функция $\phi(t)$.

На рис. 16в маркеры (о) и (х) отмечают моменты времени открытия и закрытия голосовой щели, полученные с помощью функции $\theta(t)$. На этом же рисунке маркеры (*) и (+) отмечают моменты открытия и закрытия голосовой щели, найденные по алгоритму временного анализа [2]. Сплошные вертикальные линии обозначают моменты открытия голосовой щели как моменты максимальной производной глоттограммы, а пунктиром показаны моменты закрытия голосовой щели как моменты минимальной производной глоттограммы. На рис. 16г видно, что экстремумы $\phi(t)$ также находятся в окрестности моментов T_{op} и T_{cl}. Здесь маркеры (о) и (х) отмечают минимум и максимум функции $\phi(t)$. Минимальное значение функции $\phi(t)$ равно 511 Гц, а максимальное – 2000 Гц, что близко к ожидаемым значениям формантных частот этой гласной.

5. ОБСУЖДЕНИЕ

Основной результат данной работы состоит в том, что параметры голосового источника связаны с величинами Δf_m – расстояниями между последовательными нулями и точками разрыва ФЧХ, а также с экстремумами кривых квазипериода

$$Q_1(t_c) = \overline{\Delta f_m}(t_c), \quad Q_2(t_c) = \max_m \{\Delta f_m\},$$

или их эвристического аналога $\theta(t)$ из формулы (6). В целом, полученные результаты по определению параметров t_0, T_{op}, T_{cl} голосового источника из анализа ФЧХ с помощью представленных выше методов оказываются удовлетворительными. Поскольку эти методы дают приближенные значения параметров, следует обсудить область их применимости и источники ошибок в используемом подходе.

Эксперименты с синтетическими сигналами для различных типов голосового источника и с реальными речевыми сигналами, записанными от разных дикторов, указывают, что ошибки определения параметров голосового источника зависят от формы импульса голосового источника. Эти формы в целом сильно отличаются от использованной в разд. 2 δ-образной формы, и поэтому не удивительно, что предлагаемые методы дают в определенных случаях значительные ошибки. Тем не менее, в среднем методы оказываются удовлетворительными.

Другой источник ошибок связан с особенностями стандартного кратковременного преобразования Фурье: при изменении длительности окна $w(t - t_c)$ в КПФ получаемые оценки меняются. В частности, при ее увеличении оценки становятся более устойчивыми, но увеличивается их погрешность. В численных экспериментах, предшествующих реальному анализу речи, необходимо выбирать оптимальную длительность окна КПФ.

Следующий источник ошибок обусловлен дискретизацией речевого сигнала с фиксированной частотой отсчетов. Это приводит к большой погрешности вычисления кратковременного преобразования Фурье для сигналов с большой частотой основного тона, поскольку на каждый период основного тона приходится мало отсчетов. Ошибки такого рода объясняют существование критической частоты F_0 основного тона, выше которой оценки параметров становятся ненадежными.

Наконец, ошибки возникают и из-за неточностей сопоставления глоттограмм в базе данных 3 с речевыми сигналами. Они связаны с различным расстоянием каждого диктора от микрофона. И хотя в базе 3 была выполнена некоторая средняя корректировка задержки речевого сигнала в измерениях глоттограмм, их ошибки все же присутствуют. К тому же сам принцип определения моментов открытия и закрытия голосовой щели по экстремумам глоттограмм содержит погрешности, не поддающиеся оценке [2].

Заметим, что полученные в данной работе результаты относятся к синтетическим сигналам или записям речи в заглушенной камере. Поэтому мы не учитываем в алгоритмах эффекты, связанные с шумами, реверберацией и др.

Важным результатом этой работы является исследование адекватности связи оценки $t_0 = 0.5\left(\overline{\Delta f_m}\right)^{-1}$ длительности открытой голосовой щели и экспериментальных данных. Оказалось, что погрешность оценки и ее дисперсия весьма малы (см. раздел 4.1). Длительность t_0 сама по себе представляет собой новый параметр, который, наряду с периодом основного тона, можно использовать, например, в задачах распознавания диктора. С помощью этой величины и параметра T_{cl}, для нахождения которого имеется ряд апробированных методов, можно найти параметр T_{op} по формуле $T_{op} = T_{cl} - t_0$. Однако для определения величины t_0 с помощью ФЧХ необходимо знать текущий период основного тона T_0.

Его можно найти по методике из разд. 4.2. Однако и здесь возникают некоторые проблемы. На рис. 7 видно, что выше некоторой критической частоты основного тона погрешность оценки F_0 становится отрицательной. Причина этого состоит в пропуске плохо обусловленных максимумов функции $\theta(t)$ при их поиске. В результате для высокой частоты F_0 могут быть получены ложные (заниженные) оценки, если априорно неизвестно примерное значение F_0. Однако, как упоминалось в разд. 4.2, фазовый анализ обеспечивает значительно меньшую ошибку нахождения параметров источника в диапазоне частот F_0 до 200–220 Гц по сравнению с алгоритмом [24], хотя в этом алгоритме ошибка находится в диапазоне $\pm 10\%$ и для частот выше 220 Гц. Сопоставление оценок для F_0, полученных двумя этими алгоритмами, позволяет повысить их надежность в диапазоне низких частот F_0, а также избежать ложных оценок в диапазоне высоких частот F_0.

Обращает на себя внимание значительное отличие в оценках частоты основного тона между сигналами из базы 1, синтезированными методом линейного предсказания, и сигналами из баз 2 и 3, в которых речевой сигнал генерировался искусственной физической моделью речевого тракта и собственно речевым трактом. Отличие возникает из-за использования различных источников голосового возбуждения, а также из-за того, что в синтезе методом линейного предсказания отсутствуют возмущающие факторы, которые присущи реальным речевым сигналам. Это заставляет с осторожностью относиться к выводам, полученным исключительно на базе синтезированных речевых сигналов.

Значительный разброс ошибок определения моментов T_{op} и T_{cl} на рис. 15 свидетельствует о необходимости обнаружения недостоверных оценок этих параметров. В работе [2] обнаружение подобных ошибок выполнялось путем анализа последовательности оценок T_{op} и T_{cl} на сегменте гласного. В нашей работе к рассмотрению принимались только такие оценки T_{op} и T_{cl}, которые не противоречат текущему значению периода основного тона. Тем не менее, и в таком подходе иногда наблюдаются недопустимо большие ошибки. Некоторая доля подобных ошибок может быть обнаружена или даже компенсирована в рамках фазового анализа с использованием информации о динамике функции $\phi(t)$ (см. рис. 16г). Экстремумы этой функции иногда оказываются лучше обусловленными, чем у функции $\theta(t)$. Это видно на нижнем графике в окрестности отсчетов времени 0.029, 0.056 и 0.074 с.

Ни один из известных алгоритмов анализа параметров речевого сигнала не обладает универсальностью, обеспечивающей малую погрешность независимо от вариаций речевого сигнала. Это относится и к оценке резонансных частот, и к оценке частоты основного тона, и к оценке моментов начала и конца действия голосового источника. Поэтому для каждого типа параметров необходимо совместно использовать методы, основанные на различных свойствах речевого сигнала. Рассмотрение рис. 16в еще раз подтверждает это. Для разметки речевого сегмента (нахождения параметров T_{op} и T_{cl}) имеет смысл использовать фазовый анализ совместно с другими алгоритмами, основанными на использовании других, не фазовых, свойств речевого сигнала. Из рисунка видно, что оценки фазового алгоритма и алгоритма временного анализа по [2] совпадают лишь частично, что позволяет обнаружить или исправить ошибки каждого из этих алгоритмов. Например, на интервале 0.05–0.06 с моменты закрытия голосовой щели определяются точнее, и доступны на тех сегментах, где фазовый анализ отказывает. В экспериментах с синтетическими сигналами было обнаружено, что из-за влияния начальных условий для некоторых периодов основного тона происходит такое изменение фазовых характеристик, что оценки моментов T_{op} и T_{cl} меняются местами. Это приводит к грубым ошибкам. Временной анализ нечувствителен к такому эффекту. В результате совместного анализа сегментов речи можно ожидать улучшения точности определения моментов открытия и закрытия голосового щели.

6. ЗАКЛЮЧЕНИЕ

Фазово-частотные характеристики предоставляют новую информацию о параметрах речевого сигнала, дополняющую обычный амплитудно-частотный анализ. Впервые выполнен математический анализ фазовых свойств голосового источника, на основе которого проведено обстоятельное компьютерное моделирование алгоритмов

определения длительности периода основного тона, длительности действия и моментов начала и конца импульсов голосового источника. Установлен диапазон значений частоты основного тона, в котором фазовый анализ обеспечивает приемлемую погрешность оценки этих параметров в задаче идентификации диктора. Совместный анализ речевого сигнала в фазово-частотной и амплитудно-частотной областях улучшает устойчивость и точность оценок параметров голосового источника.

При выполнении работы второй автор пользовался поддержкой Программы повышения конкурентоспособности Национального исследовательского ядерного университета МИФИ (Московского инженерно-физического института), проект № 02.a03.21.0005 от 27.08.2013.

ПРИЛОЖЕНИЕ 1. СВЯЗЬ ПАРАМЕТРА t_0 И НУЛЕЙ ФЧХ

Основываясь на соотношениях из разд. 2.1, рассмотрим случай малых потерь в речевом тракте, полагая $\delta_n = \delta = \text{const}$, $\delta/\omega_1 \ll 1$. Тогда функция $S_2(\omega)$ при $\omega \gg \omega_{N_0}$ имеет асимптотику:

$$S_2(\omega) \approx \sum_{n=1}^{N_0} \frac{\alpha_n \omega_n}{(\omega^2 - \omega_n^2)} \left(1 - \frac{i\delta}{\omega + \omega_n} - \frac{i\delta}{\omega - \omega_n}\right) \approx$$
$$\approx \sum_{n=1}^{N_0} \frac{\alpha_n \omega_n}{(\omega^2 - \omega_n^2)} \left(1 - \frac{2i\delta\omega}{\omega^2 - \omega_n^2}\right) \approx$$
$$\equiv \left(1 - \frac{2i\delta}{\omega}\right) \sum_{n=1}^{N_0} \frac{\alpha_n \omega_n}{(\omega^2 - \omega_n^2)}.$$

Отсюда

$$\ln S_2(\omega) = \ln |S_2(\omega)| + i \arg S_2(\omega) \approx \ln\left(1 - \frac{2i\delta}{\omega}\right) +$$
$$+ \ln \sum_{n=1}^{N_0} \frac{\alpha_n \omega_n}{(\omega^2 - \omega_n^2)} \approx -\frac{2i\delta}{\omega} + \ln \sum_{n=1}^{N_0} \frac{\alpha_n \omega_n}{(\omega^2 - \omega_n^2)},$$
$$\omega > \omega_1.$$

Значит, $\arg S_2(\omega) \approx -\dfrac{2i\delta}{\omega}$, и из (3) получим:

$$\varphi(\omega) \approx \left[\operatorname{Arg}\left((B\cos\omega t_0 - A) - i\left(B\sin\omega t_0 + \frac{2\delta}{\omega}\right)\right)\right]_\pi.$$

Поэтому нули и точки разрыва ФЧХ в области $\omega > \omega_1$ находятся из условия обращения в нуль величины $\operatorname{Im}\left\{(B\cos\omega t_0 - A) - i\left(B\sin\omega t_0 + \dfrac{2\delta}{\omega}\right)\right\}$, т.е. из равенства $B\sin\omega t_0 + \dfrac{2\delta}{\omega} = 0$. Можно найти асимптотическое поведение решений $\omega = \Omega_m = 2\pi f_m$ этого уравнения при $\delta \to 0$: $\Omega_m t_0 \approx \pi m + (-1)^{m-1}\dfrac{\gamma\delta}{\Omega_m}$, $\gamma = \dfrac{2}{B}$. Запишем это равенство через частоты f_m:

$$t_0\left(2\pi f_m + (-1)^m \frac{\gamma\delta}{m}\right) \approx \pi m,$$
$$t_0\left(2\pi f_{m+1} + (-1)^{m+1} \frac{\gamma\delta}{m+1}\right) \approx \pi(m+1),$$

и далее преобразуем по схеме

$$t_0\left(2\pi(\Delta f_m) + (-1)^{m+1}\gamma\delta\left(\frac{1}{m} + \frac{1}{m+1}\right)\right) \approx \pi \Rightarrow$$
$$\Rightarrow 2t_0\left(\frac{1}{N}\sum_{m=1}^{N}\Delta f_m + \frac{\gamma\delta}{\pi}\frac{1}{N}\sum_{m=1}^{N}(-1)^{m+1}\left(\frac{1}{m} + \frac{1}{m+1}\right)\right) \approx 1.$$

Из последнего равенства, переходя к пределу при $N \to \infty$, получим:

$$\overline{\Delta f_m} = \lim_{N\to\infty}\frac{1}{N}\sum_{m=1}^{N}\Delta f_m = \frac{1}{2t_0},$$

и поэтому $t_0 = \dfrac{1}{2\overline{\Delta f_m}}$.

ПРИЛОЖЕНИЕ 2. ФЧХ ДЛЯ КОНЕЧНОГО ПЕРИОДА ОСНОВНОГО ТОНА

Для источника $q(t) = \sum_{m=0}^{M-1}[A\delta(t - mT_0) - B\delta(t - t_0 - mT_0)]$, $A, B > 0$, с длительностью t_0 открытой ГЩ и длиной периода основного тона T_0 сигнал, согласно формуле (1), имеет вид

$$s(t) = \int_0^t K(t-\tau)q(\tau)d\tau =$$
$$= \sum_{m=0}^{M-1}\left[A\int_0^t K(t-\tau)\delta(\tau - mT_0)d\tau - B\int_0^t K(t-\tau)\delta(\tau - t_0 - mT_0)d\tau\right] =$$
$$= \sum_{m=0}^{M-1}[Ah(t-mT_0)K(t-mT_0) - Bh(t-t_0-mT_0)K(t-t_0-mT_0)].$$

Его преобразование Фурье вычисляется так:

$$F[s](\omega) = \sum_{m=0}^{M-1}\left[A\int_0^{+\infty} e^{-i\omega t}h(t-mT_0)K(t-mT_0)dt - B\int_0^{+\infty} e^{-i\omega t}h(t-t_0-mT_0)K(t-t_0-mT_0)dt\right] =$$
$$= \sum_{m=0}^{M-1}\left[Ae^{-i\omega mT_0}\int_0^{+\infty} e^{-i\omega\tau}K(\tau)d\tau - Be^{-i\omega(mT_0+t_0)}\int_0^{+\infty} e^{-i\omega\tau}K(\tau)d\tau\right] = \tilde{K}(\omega)\sum_{m=0}^{M-1}\left[Ae^{-i\omega mT_0} - Be^{-i\omega(mT_0+t_0)}\right] =$$
$$= \tilde{K}(\omega)\sum_{m=0}^{M-1} e^{-i\omega mT_0}\left[A - Be^{-i\omega t_0}\right].$$

Здесь $\tilde{K}(\omega) = \int_0^{+\infty} e^{-i\omega\tau}K(\tau)d\tau = \sum_{n=1}^{N_0}\frac{\alpha_n\omega_n}{(\omega+\omega_n+i\delta_n)(\omega-\omega_n+i\delta_n)}$.

Отсюда для комплексного спектра сигнала следует представление:

$$\Phi(\omega) = F[s](\omega) = \tilde{K}(\omega)\left[A - Be^{-i\omega t_0}\right]\frac{1-e^{-i\omega MT_0}}{1-e^{-i\omega T_0}}.$$

При малых потерях в тракте, полагая $\delta = \delta_n \ll \omega_1$, отсюда получим:

$$\Phi(\omega) = \left(1-\frac{2i\delta}{\omega}\right)\sum_{n=1}^{N_0}\frac{\alpha_n\omega_n}{(\omega^2-\omega_n^2)}\left[A - Be^{-i\omega t_0}\right]\frac{1-e^{-i\omega MT_0}}{1-e^{-i\omega T_0}}.$$

При больших частотах, т.е. при $\frac{\delta}{\omega} \ll 1$ и при $M = 2$ (т.е. для сигнала на одном периоде) это дает:

$$\Phi(\omega) = \left[A - Be^{-i\omega t_0}\right](1+e^{-i\omega T_0})\sum_{n=1}^{N_0}\frac{\alpha_n\omega_n}{(\omega^2-\omega_n^2)}.$$

Фаза обращается в нуль при $\mathrm{Im}\,S(\omega) = 0$. Отсюда и из последнего равенства получается уравнение $B\sin\omega t_0(1+\cos\omega T_0) = (A - B\cos\omega t_0)\sin\omega T_0$ для нулей ФЧХ. Оно аналитически не решается полностью относительно ω. Можно свести это уравнение к виду:

$$\cos\frac{\omega T_0}{2}\left(B\sin\left(\omega t_0 + \frac{\omega T_0}{2}\right) - A\sin\frac{\omega T_0}{2}\right) = 0$$

и найти явно одну серию решений $\omega_k = \frac{\pi(1+2k)}{T_0} \Rightarrow \Delta f_k = \frac{1}{T_0}$.

СПИСОК ЛИТЕРАТУРЫ

1. *Ananthapadmanabha T., Yegnanarayana B.* Epoch extraction from linear prediction residual for identification of closed glottis interval // IEEE Transactions on Acoustics, Speech and Signal Processing. 1979. V. 27. № 4. P. 309–319.
2. *Сорокин В.Н.* Сегментация периода основного тона голосового источника // Акуст. журн. 2016. Т. 62. № 2. С. 247–258.
3. *Drugman T., Thomas M., Gudnason J., Naylor P., Dutoit T.* Detection of glottal closure instants from speech signals: A quantitative review // IEEE Transactions on Audio, Speech, and Language Processing. 2012. V. 20. № 3. P. 994–1006.
4. *Sorokin V.N., Leonov A.S.* Determination of a vocal source by the spectral ratio method // Pattern Recognition and Image Analysis. 2017. V. 27. № 1. P. 139–151.
5. *Oppenheim A.V., Lim J.S.* The importance of phase in signals // Proc. IEEE. 1981. V. 69. № 5. P. 529–541.
6. *Liu L., He J., Palm G.* Effects of phase on the perception of inter-vocalic stop consonants // Speech Commun. 1997. V. 22. № 4. P. 403–417.
7. *Paliwal K.K., Alsteris L.D.* Usefulness of phase spectrum in human speech perception // Proceedings of the Eurospeech. 2003. P. 2117–2120.
8. *Laitinen M.-V., Disch S., Pulkki V.* Sensitivity of human hearing to changes in phase spectrum // J. Audio Eng. Soc. 2013. V. 61. № 11. P. 860–877.
9. *Raitio T., Juvela L., Suni A., Vainio M., Alku P.* Phase perception of the glottal excitation and its relevance in statistical parametric speech synthesis // Speech Communication. 2016. V. 81. P. 104–119
10. *Aarabi P., Shi G., Shanechi M., Rabi S.A.* Phase-Based Speech Processing. World Scientific Publishing. 2006.
11. *Mowlaee P., Saeidi R., Stylianou Y.* Advances in phase-aware signal processing in speech communication // Speech Communication. 2016. V. 81. P. 1–29.
12. *Yegnanarayana B., Sreekanth J., Rangarajan A.* Waveform estimation using group delay processing // IEEE Transactions on Audio, Speech, and Language Processing. 1985. V. 33. № 4. P. 832–836.
13. *Smits R., Yegnanarayana B.* Determination of instants of significant excitation in speech using group delay function // IEEE Transactions on Audio, Speech, and Language Processing. 1995. V. 3. № 5. P. 325–333.
14. *Brookes M., Naylor P.A., Gudnason J.* A quantitative assessment of group delay methods for identifying glottal closures in voiced speech // IEEE Trans. on Speech & Audio Processing. 2006. V. 14. № 2. P. 456–466.
15. *Cohen L.* Time-frequency distributions – a review // Proc. IEEE. 1989. V. 77. № 7. P. 941–981.
16. *Vijayan K., Kumar V., Murty K.S.R.* Feature extraction from analytic phase of speech signals for speaker verification // Speaker Odyssey. 2014. P. 1658–1662.
17. *Patterson R.D.* A pulse ribbon model of monaural phase perception // J. Acoust. Soc. Am. 1987. V. 82. № 5. P. 1560–1586.

18. *Kim D.-S.* On the perceptually irrelevant phase information in sinusoidal representation of speech // IEEE Trans. Speech Audio Process. 2001. V. 9. № 8. P. 900–905.

19. *Леонов А.С., Сорокин В.Н.* Об однозначности определения голосового источника по речевому сигналу и формантным частотам // Докл. Акад. наук. 2012. Т. 444. № 5. С. 492–495.

20. *Леонов А.С., Сорокин В.Н.* Верхняя граница ошибок решения обратной задачи определения голосового источника // Акуст. журн. 2017. Т. 63. № 5. С. 532–545.

21. CMU ARCTIC speech synthesis databases. http://festvox.org/cmu arctic

22. *Alku P., Murtola T., Malinen J., Kuortti J., Story B., Airaksinen M., Salmi M., Vilkman E., Geneid A.* OPENGLOT – An open environment for the evaluation of glottal inverse filtering // Speech Communication. 2019. V. 107. P. 38–47. https://doi.org/10.1016/j.specom.2019.01.005

23. *Fant G., Liljencrants J., Lin Q.A.* A four parameter model of glottal flow // STL–QPSR. 1985. V. 4. P. 1–13.

24. *Tsyplikhin A.I.* Analysis of vocal pulses in a speech signal // Acoust. Phys. 2007. V. 53. № 1. P. 105–118.