# Анализ гласных в фазовой области[1]

***Сорокин В. Н.,*** *доктор физико-математических наук, ведущий научный сотрудник, Институт проблем передачи информации, Российская академия наук, Москва, Россия, vns@iitp.ru*

***Леонов А. С.,*** *доктор физико-математических наук, профессор, Национальный исследовательский ядерный университет «МИФИ», Москва, Россия*

Представлены алгоритмы и численные эксперименты по определению частоты основного тона, моментов начала и конца действия голосового источника, моментов начала и конца гласных в слитном потоке речи, а также треков формантных частот. Анализ параметров речевого сигнала выполняется с использованием амплитудного динамического спектра, производной динамических фазовых спектров сигнала по частоте, а также смешанной производной фазового спектра по частоте и времени.

*• анализ речевого сигнала • групповая задержка • мгновенная частота • параметры основного тона • формантные частоты*

## ВВЕДЕНИЕ

Наиболее успешные системы речевых технологий используют статистический анализ динамического спектра мощности или его кепстрального преобразования в шкале мел. Такое представление речевого сигнала содержит информацию одновременно об амплитудах и фазах речевого сигнала. Например, системы распознавания на основе скрытых марковских моделей используют кепстральное преобразование, а нейронные сети используют либо кепстры, либо динамические спектры мощности в шкале мел.

Эффективность систем автоматического распознавания речи, основанных на описании речевого сигнала его спектром мощности, существенно ухудшается в присутствии шумов и реверберации помещения. В [1] приводятся оценки влияния этих факторов. Так, при ухудшении отношения сигнал/шум с 20 до 10 дБ, ошибка распознавания слов увеличивается с 2.6 до 30 %, т. е. примерно в 10 раз. Учитывая, что отношение сигнал/шум, равное 10 дБ, является наиболее реалистической оценкой условий распознавания в обычной акустической среде, такая ошибка практически исключает использование системы автоматического

[1] Работа А. С. Леонова поддержана Программой повышения конкурентоспособности Национального исследовательского ядерного университета МИФИ (проект 02.a03.21.0005 от 27.08.2013).

распознавания речи. Реверберация приводит к повышению ошибки распознавания при отношении сигнал/шум в 20 дБ, т. е. очень хороших акустических условиях, примерно до 20 %, а при отношении сигнал/шум в 15 дБ — до 27 %. Различие между типами микрофонов, использовавшихся для обучения и распознавания, существенно ухудшает эффективность распознавания речи и особенно диктора.

Кроме этого стандартного подхода, в поисках улучшения качества систем автоматического распознавания речи и диктора активно рассматривается возможность раздельного использования амплитудного и фазового спектра. В частности, в [2, 3] представлен обзор методов использования информации о фазе в задачах распознавания и синтеза речи, верификации диктора и предотвращения подделок речевых сигналов с помощью так называемой стеганографии (water-marking). Фазовые параметры используются в [4] для верификации диктора, в [5] для детектирования синтетического сигнала при верификации диктора, а в [6] — в задаче разделения голосов одновременно говорящих дикторов. В [7] фаза сигнала на нулевой частоте применяется для детектирования сегмента с голосовым возбуждением.

Другой способ формирования исходных данных для распознавания речи и диктора, отличный от статистического, состоит в сегментации речевого сигнала на фонетически значимые элементы и описания этих элементов в терминах специфических признаков. Исторически этот подход развивался на первых этапах исследования проблемы автоматического распознавания, но оказался трудно реализуемым в силу разнообразных источников изменчивости параметров речевого сигнала и малой мощности вычислительных средств. К настоящему времени накоплены критически важные знания о свойствах речевого сигнала, которые позволяют вернуться к реализации этого подхода. В частности, исследование фазовых параметров открывает новые возможности для детектирования элементов речевого кода, сегментации периода основного тона на участки с открытой и закрытой голосовой щелью, оценки формантных частот.

Гласные являются важным элементом речевого кода на его нижнем уровне, который включает также звонкие и глухие смычки, назальные и фрикативные звуки. Формантные частоты гласных используются для распознавания, а также в обратных задачах, таких как определение формы речевого тракта и вычисление команд управления артикуляцией. В обычной речи гласные характеризуются присутствием голосового источника, формантной структурой и относительно высокой энергией. Моменты открытия и закрытия голосовой щели в фазовой области исследовались в [8-12], а сегментация гласных на основе агрегирования автокорреляционных коэффициентов различных фазовых параметров была описана в [12]. В [12, 13] также рассматривалась возможность оценки формантных частот по фазовым параметрам.

В данной работе предпринимается попытка разработки алгоритмов сегментации речевого сигнала на гласно-подобные звуки и анализа гласных звуков, включая оценку частоты основного тона и формантных частот путем агрегирования фазовых параметров и параметров динамического спектра амплитуд.

## 1. МАТЕМАТИЧЕСКИЕ МОДЕЛИ ПАРАМЕТРОВ ФАЗОВОЙ ФУНКЦИИ

Комплексный спектр $S(\omega,t)$ некоторого сигнала можно представить двояко:

$$S(\omega,t) = A(\omega,t)e^{j\varphi(\omega,t)} = u(\omega,t) + jv(\omega,t)\,,$$

где $A(\omega,t)$ — амплитудный динамический спектр, а $\varphi(\omega,t)$ — фазовый динамический спектр, и

$$A(\omega,t)=\left|S(\omega,t)\right|=[u^2(\omega,t)+v^2(\omega,t)]^{1/2}.$$

Мнимая и действительная компоненты комплексного спектра определяются через амплитуду и фазу как

$$u(\omega,t)=A(\omega,t)\cos\varphi(\omega,t)$$
$$v(\omega,t)=A(\omega,t)\sin\varphi(\omega,t).$$

Амплитудный динамический спектр $A(\omega,t)$ удобно представлять графически в виде сонограммы (см., например, рис. 1а). Это позволяет визуально идентифицировать различные характеристики, связанные с акустикой или артикуляцией речевого сигнала. Можно попытаться представить фазовый динамический спектр $\varphi(\omega,t)$ графически аналогичным образом. При этом надо учитывать, что функция $\varphi(\omega,t)$ при каждом фиксированном $t$ имеет область значений $[-\pi,\pi]$ и является, вообще говоря, разрывной функцией аргумента $\omega$. Кроме того, существуют различные способы преобразования разрывной фазы $\varphi(\omega,t)$ в непрерывную функцию частоты, которую также можно изобразить по аналогии с сонограммой. Однако оба этих изображения (фазограммы) не демонстрируют никаких структурных особенностей, характерных для сонограмм. Поэтому, используя их, практически невозможно найти какие-либо специфические структуры на интервале речевого сигнала. Это можно проиллюстрировать сонограммой и изображениями разрывной и непрерывной фазовой функции $\varphi(\omega,t)$ для слова/*восемь* (Рис. 1).

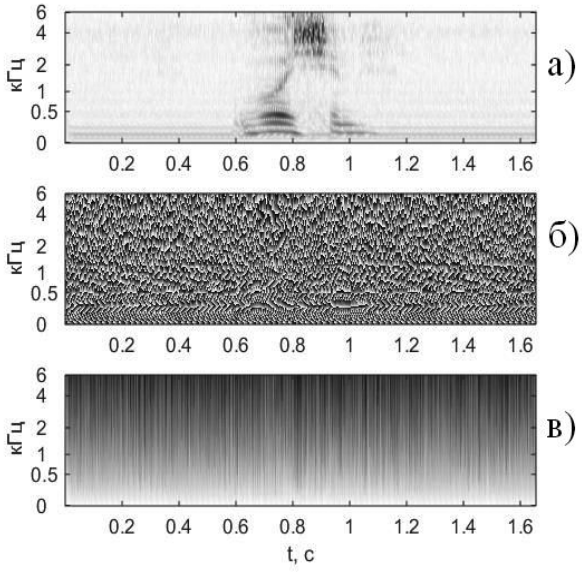


*Рис. 1.* Сонограмма слова/восемь/(а); фазограмма разрывной фазовой функции (б); фазограмма непрерывной фазовой функции (в).

В [12] разрывная фаза использована для оценки моментов начала и конца действия голосового источника. Однако обнаружить структуры,

коррелированные с акустическими признаками фонетических элементов в речевом сигнале, в соответствующих фазограммах не удается. Оказалось, что это возможно лишь в фазограммах производных непрерывной фазовой функции по частоте и времени: $\varphi_\omega = \partial\varphi(\omega,t)/\partial\omega$ и $\varphi_t = \partial\varphi(\omega,t)/\partial t$. Эти производные можно выразить через действительную и мнимую часть комплексного спектра сигнала следующим образом:

$$\varphi_\omega = \frac{u(\omega,t)v'(\omega,t) - v(\omega,t)u'(\omega,t)}{u^2(\omega,t) + v^2(\omega,t)} = \frac{u(\omega,t))v'(\omega,t) - v(\omega,t)u'(\omega,t)}{\left|S(\omega,t)\right|^2},$$

$$\varphi_t = \frac{u(\omega,t)\dot{v}(\omega,t) - v(\omega,t)\dot{u}(\omega,t)}{u^2(\omega,t) + v^2(\omega,t)} = \frac{u(\omega,t)\dot{v}(\omega,t) - v(\omega,t)\dot{u}(\omega,t)}{\left|S(\omega,t)\right|^2}.$$

Здесь штрих обозначает производную по частоте, а точка — производную по времени. Эти формулы позволяют найти более употребительные в речевом анализе величины: групповую задержку $\tau(\omega,t) = -\varphi_t/2\pi$ и мгновенную частоту $Q(\omega,t) = \varphi_t/2\pi$. В [12] было установлено, что полезными для анализа свойствами обладает также смешанная производная фазы по частоте и времени

$$\varphi_{\omega t} = \frac{\partial^2\varphi(\omega,t)}{\partial\omega\partial t} = \frac{\partial}{\partial\omega}\left[\frac{u(\omega,t)\dot{v}(\omega,t) - \dot{u}(\omega,t)v(\omega,t)}{u^2(\omega,t) + v^2(\omega,t)}\right].$$

Для иллюстрации на рисунке 2 для речевого сигнала слова/*восемь*/в разные моменты времени (0.3 с и 0.75 с) представлены как функции частоты значения разрывной и непрерывной фазы, групповой задержки и мгновенной частоты. Видно, что непрерывная фаза содержит значительную линейную компоненту. Поэтому ее можно представить в виде суммы двух членов — линейной функции частоты $\varphi_L(\omega,t) = \varphi_0 + k_\varphi\omega$ и нелинейной $\varphi_A(\omega,t)$, т.е. $\varphi(\omega,t) = \varphi_L(\omega,t) + \varphi_A(\omega,t)$. Как будет показано ниже, модуляции коэффициента $k_\varphi(t)$ важны для оценки параметров периода основного тона и сегментации гласных.

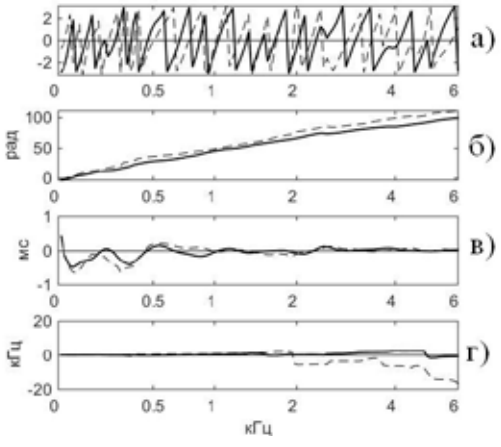


*Рис. 2.* Разрывная фаза (а), непрерывная фаза (б), групповая задержка (в), мгновенная частота (г). Отсчет на паузе (---) 0.3 сек, отсчет на гласном (–) 0.75 сек

Анализ комплексного спектра одиночного осциллятора показывает, что максимум производной $\varphi_\omega$ его фазовой функции по частоте приходится на собственную частоту осциллятора. Однако в спектре функции $\varphi_\omega$ для суммарного сигнала от нескольких осцилляторов, наряду с пиками на собственных частотах, появляются отрицательные экстремумы [12]. В этой же работе было установлено, что для одиночного осциллятора в спектре производной фазы по времени $\varphi_t$ присутствуют пики не только на собственной частоте осциллятора, но и на гармониках этой частоты. При этом спектр сигнала от суммы осцилляторов содержит множество экстремумов, среди которых практически невозможно определить собственные частоты этих осцилляторов. Как показали эксперименты с реальными речевыми сигналами, при оценке формантных частот вместо производной фазы по времени $\varphi_t$ целесообразно использовать смешанную производную $\varphi_{\omega t}$.

Комплексный спектр сигнала действительной переменной *s* (*t*) обычно вычисляют с помощью кратковременного преобразования Фурье в скользящем окне *w:*

$$S(\omega,t)=\int_0^\infty w(t-\tau)s(\tau)e^{-j\omega\tau}d\tau\ .$$

Ширина и вид окна *w* существенно влияют на комплексный спектр и в каждом конкретном случае эти параметры подбираются экспериментально. Далее анализ речевого сигнала выполняется на основе динамического спектра, представленного как $A(\omega,t)=\log\left|S(\omega,t)\right|^2$.

Другой способ получения спектра состоит в использовании гребенки из *N* фильтров, где амплитуда $A_k(t)$ вычисляется как огибающая сигналов $s_k(t), k=1,2,...N$ на выходе каждого фильтра. Обычно применяют фильтры Габора, чье преимущество состоит в линейности фазы. В применении к анализу речевого сигнала представляется целесообразным использовать фильтры, моделирующие свойства периферического отдела слухового анализа. К числу таких фильтров относится система так называемых гамма-тон фильтров [14].

Амплитудную и фазовую компоненты комплексного спектра можно получить, используя вместо преобразования Фурье преобразование Гильберта. Применяя преобразование Гильберта $H\{s_k(t)\}$ к сигналу $s_k(t)$, получим так называемые аналитические сигналы $x_k(t)=s_k(t)+jH\{s_k(t)\}, k=1,2,...$ Затем из представлений $x_k(t)=A_k(\omega,t)e^{j\varphi_k(\omega,t)}$ находятся дискретные по частоте и времени компоненты $A_k(\omega,t)$, $\varphi_k(\omega,t)$ комплексного спектра.

Техническое преимущество использования аналитического сигнала $x_k(t)$ вместо $s_k(t)$ состоит в возможности его дискретизации с вдвое меньшей частотой Найквиста-Котельникова, что ускоряет его обработку. Другое отличие заключается в том, что применение преобразования Гильберта вместо кратковременного преобразования Фурье не требует выбора формы и длительности окна $w$, поскольку преобразование Гильберта выполняется на всем интервале существования сигнала.

Получив тем или иным способом комплексный спектр регистрируемого микрофоном речевого сигнала, его далее используют для речевого анализа. В этом спектре речевые компоненты отягощены различными видами искажений. В определенном приближении можно считать, что спектр состоит из нескольких компонент и представим в виде:

$$S(\omega,t) = S_{mic}(\omega,t)[S_{rad}(\omega,t) + S_{rv}(\omega,t) + S_N(\omega,t)],$$

где $S_{mic}(\omega,t)$ — передаточная функция микрофона; $S_{rad}(\omega,t)$ — спектр сигнала, излучаемого из рта диктора; $S_{rv}(\omega,t)$ — суммарный сигнал от множества отражений излученного сигнала от поверхностей помещения, $S_{rv}(\omega,t) = f\{S_{vt}(\omega,t)\}$ ; $S_N(\omega,t)$ — спектр аддитивного шума среды и наводки электрических сетей. В свою очередь, спектр истинного речевого сигнала в предположении линейности взаимодействия голосового источника и речевого тракта представляется как $S_{rad}(\omega,t) = S_{es}(\omega,t)S_{vt}(\omega,t)$. Здесь $S_{es}(\omega,t)$ — спектр источника возбуждения акустических колебаний в речевом тракте, $S_{vt}(\omega,t)$ — передаточная функция речевого тракта. Если расстояние и направление диктора на микрофон не изменяются в процессе разговора, то передаточная функция микрофона постоянна во времени, и зависит только от частоты $S_{mic}(\omega,t) = S_{mic}(\omega)$. Тогда в производной по времени логарифма спектра $\partial[\log(S(\omega,t)]/\partial t$ отсутствует влияние характеристик микрофона.

Динамика компонент комплексного спектра $S_{rad}(\omega,t)$, т. е. соответствующих амплитудных и фазовых спектров, по-разному зависит от процессов речеобразования. Так, спектр $S_{es}(\omega,t)$ отображает относительно быстро меняющиеся характеристики голосового источника, а $S_{vt}(\omega,t)$ связан с медленными артикуляторными движениями. Поэтому их можно разделить, фильтруя или сглаживая на интервалах разной длительности, например, 2.5 мс и 25 мс. В [12] было установлено, что кратковременные модуляции параметров фазовой функции содержат информацию о параметрах голосового источника, таких как: длительность открытой голосовой щели, период основного тона, моменты открытия и закрытия голосовой щели. Долговременные модуляции этих параметров отражают медленные артикуляторные движения. Формантные частоты могут быть определены как по кратковременным, так и долговременным модуляциям. Различие в кратковременных и долговременных модуляциях иллюстрируется рисунком 3.

## 2. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

Анализ фазовых параметров выполнялся с использованием гребенки на 128 гамма-тон фильтров в частотной шкале мел и предискажением с подъемом высоких частот.

### 2.1. Сегментация гласных

Сегментация гласных с хорошей точностью при ручной разметке в спектрально-временной области описана в [15, 16]. Однако предлагаемый там алгоритм требует доработки с целью повышения его устойчивости к шумам и искажениям при автоматической разметке. Проанализируем процедуры сегментации и возможности их улучшения с использованием фазовых спектров.

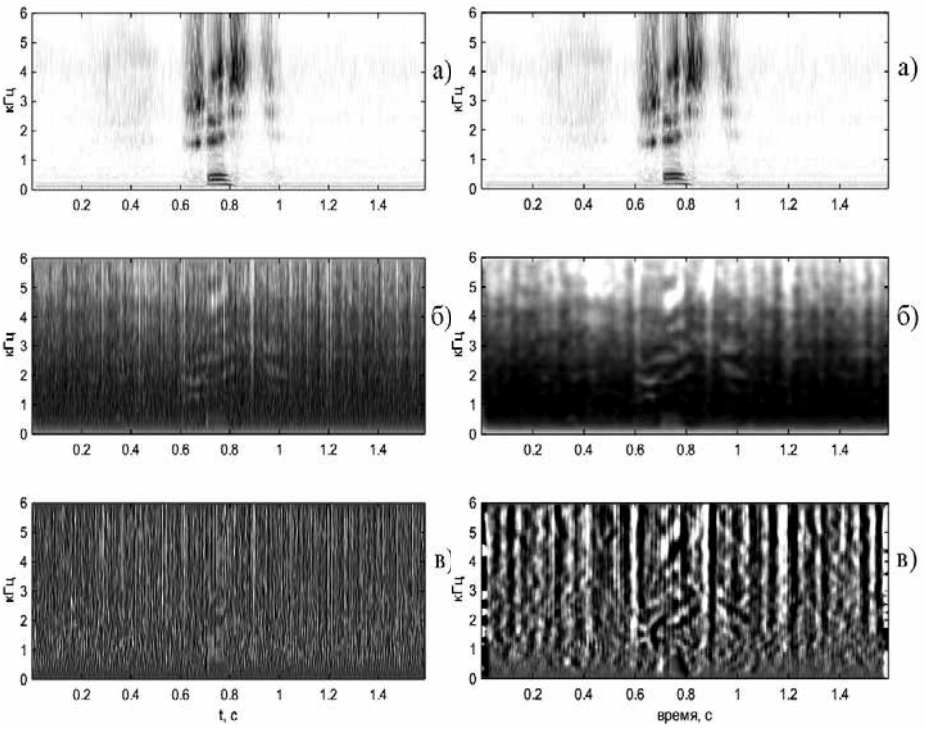


*Рис. 3.* Сонограмма слова/*шесть*/(а), групповая задержка (б), смешанная производная (в). Слева — кратковременные модуляции со сглаживанием на интервале 2.5 мс; справа — долговременные модуляции со сглаживанием на интервале 25 мс.

Одним из важных признаков гласных в обычной (не шепотной) речи является присутствие голосового возбуждения. Помимо гласных, голосовой источник участвует также в производстве других фонетических элементов, таких, как /*в, л, р*/, а также звонкие смычки взрывных согласных/ *б, д, г*/, назальных/*м, н*/и звонких фрикативных/*з, ж*/. Определение начала и конца сегмента речи с голосовым возбуждением требуется при формировании детекторов артикуляторных событий [17]. При этом необходимо по возможности исключить участки речевого сигнала со звонкими смычками. Обычно присутствие голосового источника детектируется путем оценки степени периодичности речевого сигнала, в частности, автокорреляционным методом. Но такой подход не позволяет различить сегменты гласных от звонких смычек.

Излучение акустических колебаний из речевого тракта при звонких смычках происходит через щеки и шею. Частота этого излучения равна частоте радиального резонанса речевого тракта, который находится в диапазоне 100-300 Гц [18], так что диапазоны частот основного тона и частот радиального резонанса перекрываются. Один из подходов к детектированию звонкой смычки состоит в вычислении отношения энергии амплитудно-частотного спектра в области радиального резонанса к средней энергии спектра в области частот выше 300 Гц. Другой способ дифференциации сегмента гласного от звонкой смычки использует оценку периодичности речевого сигнала в диапазоне частот от первой до второй форманты, т. е. 1000-3000 Гц [19].

В дополнение к этим методам целесообразно использовать свойства фазовой функции речевого сигнала. Ниже описывается алгоритм сегментации гласных с использованием параметров фазовой функции. На первом шаге вычисляются коэффициенты автокорреляции следующих параметров фазовой функции, сглаженных на интервале 2.5 мс:

1. Производная по времени от коэффициента линейной компоненты фазы $\dot{k}_{\varphi}(t)$.
2. Производная по времени функции $\dot{\theta}(t)$, пропорциональной максимальному интервалу между нулями разрывной фазовой функции.
3. Производная по времени от модуляции средней мгновенной частоты в диапазоне частот 1000-3000 Гц $\dot{\bar{Q}}(\omega,t)/\omega$.
4. Производная по времени от модуляции средней энергии амплитудного спектра в диапазоне частот 1000-3000 Гц $\dot{\bar{A}}(t)$.

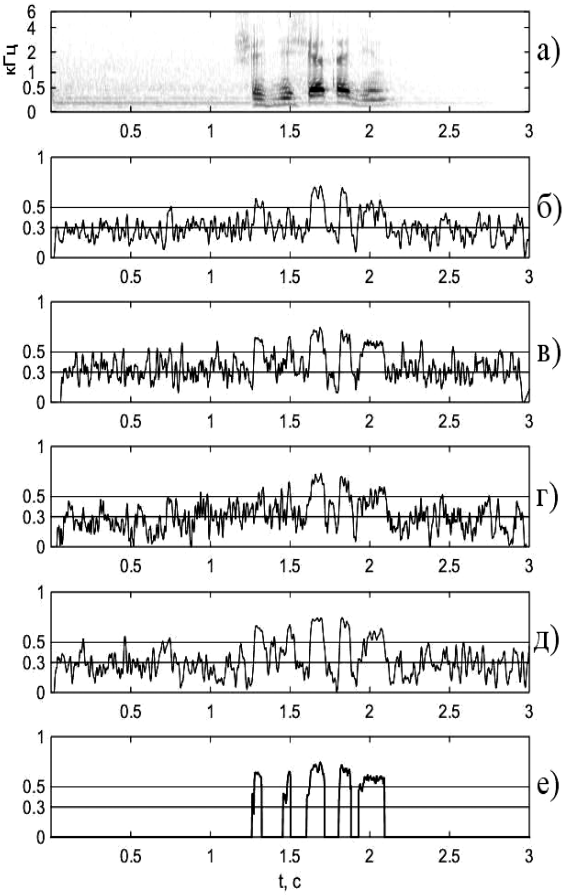


*Рис. 4.* Сонограмма (а) и коэффициенты автокорреляции функций $\dot{k}_{\varphi}(t)$ (б), $\dot{\theta}(t)$ (в), $\dot{\bar{Q}}(\omega,t)/\omega$ (г), $\dot{\bar{A}}(t)$ (д), результирующая оценка коэффициента автокорреляции (е)

Далее коэффициенты автокорреляции этих параметров агрегируются путем выбора среди этих параметров в каждый момент времени максимального коэффициента и соответствующего ему значения частоты основного тона $F_0$. Отсев ложных срабатываний выполняется с использованием ограничений на минимальную длительность сегмента с высоким коэффициентом автокорреляции (40 мс) и разность соседних оценок частоты основного тона (с порогом $0.2F_0$). Текущее значение периода основного тона $T_0$ определяется как разность между соседними минимумами вышеупомянутых параметров.

Этот алгоритм был использован в ряде численных экспериментов. На рис. 4 показаны коэффициенты автокорреляции перечисленных функций, а также результирующие значения агрегированного коэффициента автокорреляции, определяющие сегментацию гласных в слитном произнесении словосочетания 61 (*шестьдесятодин*).

### 2.2. Оценка и сегментация периода основного тона

Оценка интервала времени между импульсами голосового источника, т. е. мгновенного периода основного тона $T_0$, необходима при просодическом анализе

для определения типа фразы (повествовательная, вопросительная и т.д.), установления места ударения в слове и логического ударения. Распределение частоты основного тона характеризует пол диктора в задаче распознавания речи, а также самого диктора в задаче идентификации и верификации. Модуляции периода основного тона могут использоваться для диагностики заболеваний гортани и оценки физического и эмоционального состояния человека. Моменты начала и конца импульса голосового возбуждения указывают на интервалы времени, которые обеспечивают наиболее точную оценку формантных частот.

Существует ряд методов оценки частоты основного тона в спектральной и временной области. В спектральной области $F_0$ оценивается по положению первого пика амплитудного спектра и разнице между гармониками основного тона. Наиболее популярны алгоритмы анализа сигнала-остатка по линейному предсказанию речевого сигнала и автокорреляционные алгоритмы. Для любых алгоритмов анализа речевого сигнала характерна определенная неустойчивость, не позволяющая полностью доверять результатам анализа. Поэтому целесообразно использовать совместную оценку речевых параметров по результатам работы алгоритмов, основанных на разных свойствах речевого сигнала. Анализ частоты основного тона в фазовой области предлагает еще одну оценку $F_0$ в дополнение к существующим методам.

Эксперименты с использованием синтезированных и реальных речевых сигналов показали, что все параметры фазовой функции отображают деятельность голосового источника. Некоторые из этих параметров можно использовать, агрегируя оценки $F_0$ по каждому параметру. Один из способов агрегирования оценок $F_0$ по параметрам, указанным в предыдущем разделе, состоит в выборе на каждом скользящем интервале времени такой $F_0$, которой соответствует наибольший коэффициент корреляции. На рисунке 5б сопоставляются результаты оценок $F_0$ с помощью автокорреляционного алгоритма из [20] и агрегированными оценками. Это делается на примере речевого сигнала для словосочетания (*шестьдесятодин*).

Фонетическая разметка этого словосочетания в терминах системы, описанной в [21], есть /Шы3`Д`Дh`иС`ЯТТ!аД`Д`!ИНН!/. Здесь заглавные буквы обозначают ударные гласные, апостроф — смягчение,! — неаспиративный взрыв смычки, а h — аспиративный взрыв. Необходимо обратить внимание на то, что в данной реализации этого словосочетания первые сегменты отличаются от буквенного написания. При этом звонкий фрикативный З' и звонкие смычки Д' игнорируются детектором гласных, тогда как назальный М опознается как сегмент с голосовым возбуждением. В других произнесениях этого словосочетания звонкий фрикативный З' может опознаваться как сегмент с голосовым возбуждением.

Как видно на рисунке 5б, оценка $F_0$ по полному речевому сигналу автокорреляционным методом и агрегированная оценка по фазовым параметрам на сегментах гласных звуков близки, хотя временами и не совпадают. В свою очередь, эти оценки также могут быть агрегированы.

Ш ы З` Д` и С` Я Т а Д` И Н

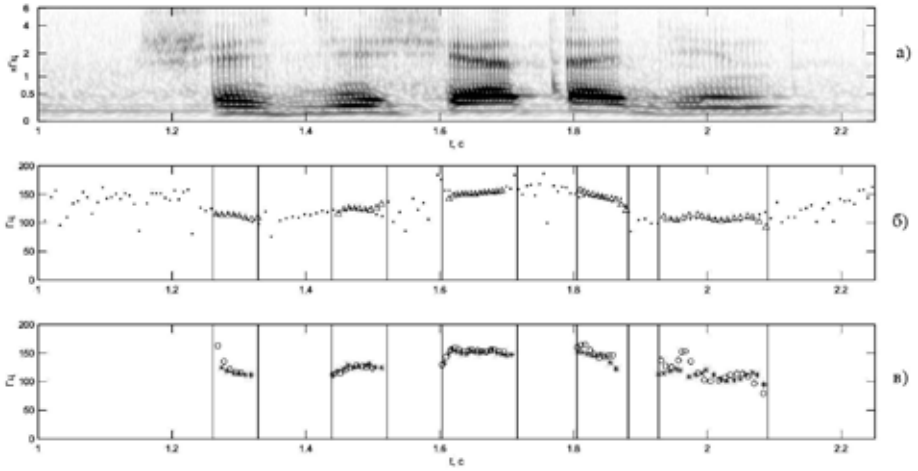


*Рис. 5.* Сонограмма (а), оценки по всему сигналу (•••) и оценки по максимуму автокорреляции (ΔΔΔ) (б), оценки в момент начала голосового возбуждения $T_{op}$ (○○○) и оценки в момент конца голосового возбуждения $T_{cl}$ (***) (в). Вертикальные линии соответствуют границам результирующей оценки коэффициента автокорреляции, показанным на рис. 4.

При этом $F_0$ по полному речевому сигналу оценивается и на звонких смычках/д/, что служит дополнительным параметром в распознавании звонких смычек. Но такая оценка осуществляется и на глухой смычке, и на паузах, причем эта оценка находится в диапазоне возможных частот $F_0$, так что ее нельзя отсеять по признаку выхода за этот диапазон.

Другой способ агрегирования оценок $F_0$ по вышеуказанным параметрам используется для определения моментов начала $T_{op}$ и конца $T_{cl}$ действия голосового источника [12]. Предполагается, что $T_{op}$ соответствует моментам максимального значения параметров, а $T_{cl}$ — моментам их минимального значения. Поскольку эти экстремумы у разных параметров подвержены различным сдвигам, то для каждой группы оценок $T_{op}$ и $T_{cl}$ выбирается их минимальное значение, т. е. минимальная задержка. Этот способ агрегирования был апробирован на синтезированных гласных и показал хорошее соответствие с известными моментами начала и конца действия голосового источника. После того, как получены оценки $T_{op}$ и $T_{cl}$ для последовательности импульсов голосового возбуждения, можно найти и соответствующие им оценки периода основного тона для каждого *k*-го импульса как $T_{0op}(k) = T_{op}(k+1) - T_{op}(k)$ и $T_{0cl}(k) = T_{cl}(k + T_{op}(k+1) - T_{cl}(k)$. Соответствующие частоты основного тона показаны на рисунке 5 в.

Этот и другие численные эксперименты показывают, что по фазовым параметрам можно получить адекватные оценки частоты основного тона.

## 2.3. Формантный анализ

### 2.3.1. Эксперименты с синтезированными звуками.

Некоторое представление о свойствах фазовых параметров при оценке формантных частот можно получить в экспериментах с синтезированными звуками с известными формантными частотами. В данной работе использовались изолированные гласные */а, э, о, у, и, ы/*, помещенные между двумя сегментами с нулевой амплитудой сигнала и длительностью 0.5 с. Затем к каждому такому сигналу добавлялась случайная функция с амплитудой, равной 0.001 от среднеквадратической амплитуды гласного. Резонансные частоты этих гласных (таблица 1) были ранее найдены путем решения волнового уравнения для соответствующей формы речевого тракта [22].

*Таблица 1*

**Резонансные частоты синтезированных гласных, Гц**

| | А | Э | О | У | И | Ы |
|---|---|---|---|---|---|---|
| F1 | 600 | 565 | 497 | 309 | 290 | 286 |
| F2 | 1200 | 1381 | 914 | 758 | 2272 | 1874 |
| F3 | 2300 | 2252 | 2316 | 2042 | 3100 | 2575 |
| F4 | 3500 | 2789 | 2625 | 2761 | 4000 | 3732 |
| F5 | 3806 | 3683 | 4030 | 3612 | 5055 | 4421 |
| F6 | 4742 | 4235 | 4728 | 4434 | 6109 | 6109 |

Были выполнены оценки формантных частот по пикам производной фазы по частоте и смешанной производной по частоте и времени, сглаженных на скользящих интервалах 2.5 мс и 25 мс. В таблицах 2, 3 показаны ошибки, не превышающие 15 % при оценке формантных частот по пикам спектров производной фазы по частоте, сглаженных на интервале 2.5 мс и 25 мс. При этом к оценке допускались лишь те пики, частота которых была наиболее близка к известным резонансным частотам соответствующего гласного.

*Таблица 2*

**Ошибки оценок формантных частот для производной фазы по частоте, %**

| | А | Э | О | У | И | Ы |
|---|---|---|---|---|---|---|
| Интервал сглаживания 2.5 мс | | | | | | |
| F1 | -13.1 | -8.2 | - | -3.2 | 14.3 | 4.5 |
| F2 | -0.4 | -7.7 | 0.1 | -1.4 | -0.7 | -2.9 |
| F3 | - | - | - | - | -2.4 | -1.7 |
| F4 | 0.5 | 1.7 | -2.2 | -1.9 | -3.2 | -3 |
| F5 | - | -1.7 | -0.9 | -0.5 | - | - |
| Интервал сглаживания 25 мс | | | | | | |
| F1 | -10.5 | - | - | 14.6 | - | - |
| F2 | - | - | - | - | -0.7 | -0.5 |
| F3 | -1.9 | - | -4.4 | -1.8 | -2.4 | -1.7 |
| F4 | -1.6 | -1 | -3.6 | -1.9 | 4.2 | -0.5 |
| F5 | - | -0.3 | -0.2 | -0.5 | -2.2 | -1.6 |

В процессе получения оценок выявлено, что популярный метод оценки формантных частот по пикам амплитудного спектра в ряде случаев приводит к существенным ошибкам (таблица 3). Здесь лишь несколько ошибок оказываются менее 5 %, четыре из 18 оценок меньше 10 %, и три оценки весьма велики.

*Таблица 3*

**Ошибки оценок формантных частот по сонограмме, %**

| | А | Э | О | У | И | Ы |
|---|---|---|---|---|---|---|
| F1 | -0.83 | -6.3 | 14.9 | -4.7 | 9 | 7.7 |
| F2 | 0.42 | 5.9 | 1.45 | -16.4 | -1.6 | 0.5 |
| F3 | -0.39 | -0.17 | 25.2 | 22.1 | 0.26 | -0.54 |

В экспериментах с синтезированными гласными также выяснилось, что, помимо треков пиков производной по частоте и смешанной производной, близких к формантам, возникают «паразитные треки», далекие от формант. Часто они образуют пары, между которыми находятся форманты, но расстояние от каждого трека до такой форманты может быть относительно велико. На рисунке 6 представлены треки пиков производной фазы по частоте и смешанной производной с оценками формантных частот в сопоставлении с истинными значениями резонансных частот гласного /*а* /в линейной шкале частот.

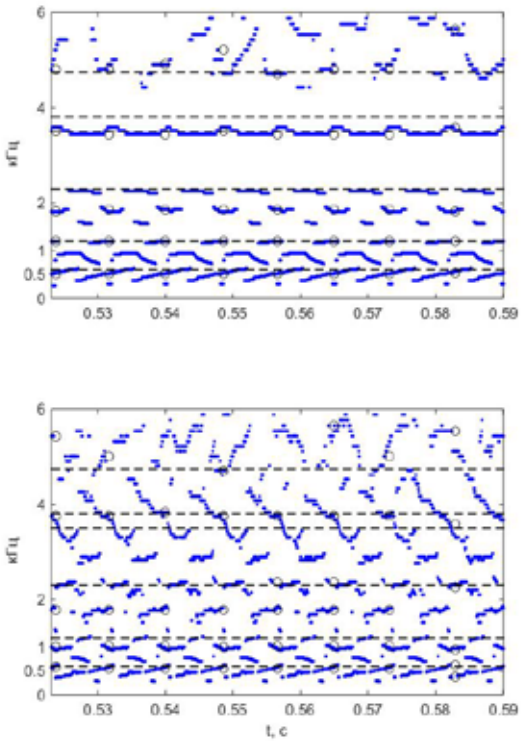


*Рис. 6.* Треки пиков производной фазы по частоте (вверху) и смешанной производной (внизу). Оценки формантных частот (○○○); истинные значения резонансных частот (---).

В целом эксперименты с синтезированными гласными показывают, что в ряде случаев возможно оценить формантные частоты с приемлемой погрешностью, если примерно известны диапазоны формант для каждого гласного. Иногда оценки

формант по фазовым параметрам выполняются даже с меньшей погрешностью, чем по пикам сонограммы. Например, это можно видеть в таблице 3 для третьей форманты. При этом обнаружено, что даже такой малый шум, который использовался в экспериментах, приводит к заметной разнице в оценке формант при повторных вычислениях.

### 2.3.2. Эксперименты с реальными сигналами.

Рассмотрим теперь динамические спектры производной фазы по частоте и смешанной производной. На рисунках 7-10 показаны фазограммы производной по частоте и смешанной производной, сглаженные на интервалах 2.5 мс и 25 мс, треки пиков этих функций, а также оценки формантных частот на сегменте гласного /э/ в слове /*шесть*/. Перед началом этого слова находится пауза длительностью 0.6 сек, а после него — длительностью 0.5 сек. Формантные частоты в начале гласного были определены вручную по пикам сонограммы. Они равны 0.4012, 1.6443, 2.2772, 2.9327, 3.8820, 5.0347 кГц. На рисунках 8 и 10 эти частоты представлены прерывистыми линиями.

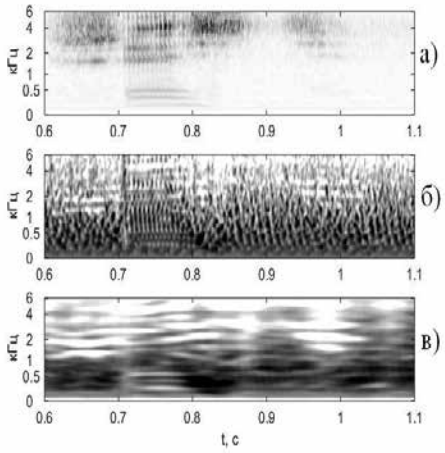


*Рис. 7.* Производная по частоте. Сонограмма (а), сглаживание на интервале 2.5 мс (б), сглаживание на интервале 25 мс (в).

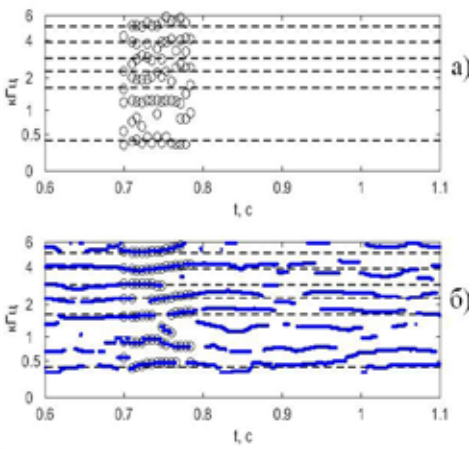


*Рис. 8.* Производная по частоте. Оценки формантных частот (○○○). Шкала мел. Сглаживание на интервале 2.5 мс (а). Треки пиков производной по частоте, сглаженных на интервале 25 мс (б).

Как видно из рисунков 8 и 10, пики производной фазы по частоте и смешанной производной на сегменте гласного не обязательно соответствуют формантным частотам. Поэтому примерная оценка формантных частот должна быть получена еще каким-то другим способом. Например, используем метод оценки формантных частот по пикам амплитудно-частотного спектра из работы [12]. Как предлагается в этом методе, треки пиков сонограммы сортируются по убыванию амплитуды при условии, что в каждый момент времени оценивается шесть пиков, а формантные частоты отсчитываются в моменты времени, несколько раньше моментов начала действия голосового источника.

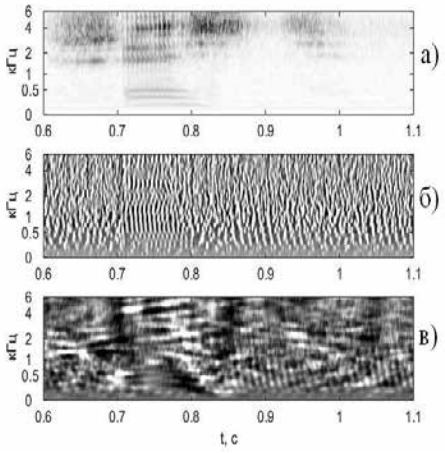


*Рис. 9.* Смешанная производная. Сонограмма (а), сглаживание на интервале 2.5 мс (б), сглаживание на интервале 25 мс (в).

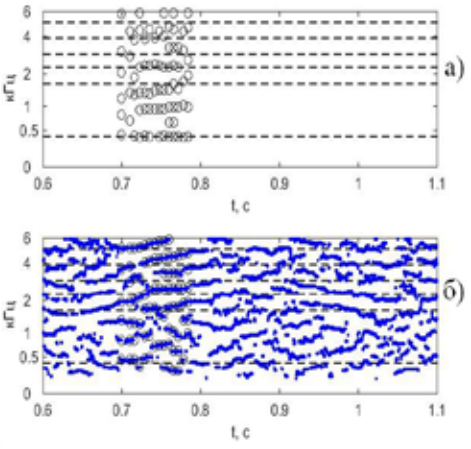


*Рис. 10.* Смешанная производная. Оценки формантных частот (○○○). Шкала мел. Сглаживание на интервале 2.5 мс (а). Треки пиков производной по частоте, сглаженных на интервале 25 мс (б).

В этом, как и в любом другом методе, также нет гарантии определения «истинных» формантных частот, поскольку некоторые оценки могут отсутствовать, и, наоборот, могут появиться лишние оценки. Коррекция оценок на этом этапе может быть выполнена с использованием сведений о распределении формантных частот по большому числу дикторов и гласных в различных контекстах. В [21] приводится 8 кластеров гласных с соответствующими весами (табл. 4).

*Таблица 4*

**Частоты центров кластеров формант (Гц) и веса кластеров**

| № кластера | вес | F1 | F2 | F3 |
|---|---|---|---|---|
| 1 | 0.365 | 317 | 1842 | 2475 |
| 2 | 0.181 | 392 | 1262 | 2339 |
| 3 | 0.138 | 342 | 907 | 2471 |
| 4 | 0.103 | 403 | 1576 | 2373 |
| 5 | 0.082 | 492 | 1124 | 2239 |
| 6 | 0.071 | 297 | 2082 | 3287 |
| 7 | 0.032 | 423 | 1076 | 3273 |
| 8 | 0.029 | 377 | 1827 | 2690 |

Аналогичная коррекция выполнена в данной работе путем поиска такой комбинации этих частот, которая обеспечивает минимальное расстояние в некоторой метрике к какому-то центру кластеров. При этом осуществляется перебор по всем возможным сочетаниям измеренных формантных частот в диапазоне до максимального значения частоты третьей форманты около 3600 Гц. Так для каждого $i$-го импульса голосового источника получается вектор формантных частот $\mathbf{F}_i = \{F_{1i}, F_{2i}, F_{3i}\}$. На следующем этапе формантные частоты по фазовым параметрам сравниваются уже с этим вектором, и для каждой частоты $F_{ki}, k = 1,2,3,$ выполняется поиск такой частоты пика фазового параметра, относительное расстояние до которого не превышает некоторой величины.

Далее необходимо использовать какой-либо способ агрегирования оценок формантных частот, полученных разными методами. Рассмотрим два алгоритма агрегирования. Пусть $t_m$ — времена, при которых оцениваются формантные частоты, и для каждого такого времени $t_m$ имеется несколько оценок форманты $F$ разными методами:

$$F^{(1)}(t_m), F^{(2)}(t_m), ..., F^{(N_m)}(t_m) = \left\{F^{(k)}(t_m)\right\}_{k=1}^{N_m}.$$

Тогда по методу наименьших квадратов (МНК) можно получать различные «скользящие» по времени осредненные оценки форманты в каждый момент $t_m$, используя данные

$$\left\{F^{(k)}(t_{m-1})\right\}_{k=1}^{N_{m-1}}, \left\{F^{(k)}(t_m)\right\}_{k=1}^{N_m}, \left\{F^{(k)}(t_{m+1})\right\}_{k=1}^{N_{m+1}}$$

для моментов $t_{m-1}, t_m, t_{m+1}$ $(m = 1, ..., M-1)$. Приведем два примера таких оценок.

Предположим, что формантная частота мало меняется на каждом отрезке $[t_{m-1}, t_{m+1}]$. Тогда на таких отрезках можно постулировать следующую зависимость форманты от времени: $F(t) = b = const$. Константа $b$ на каждом отрезке может быть разной. При таком предположении функционал МНК имеет вид

$$\Phi(b) = \sum_{k=1}^{N_{m-1}} \left(b - F^{(k)}(t_{m-1})\right)^2 + \sum_{k=1}^{N_m} \left(b - F^{(k)}(t_m)\right)^2 + \sum_{k=1}^{N_{m+1}} \left(b - F^{(k)}(t_{m+1})\right)^2.$$

Из условия его минимума $\Phi'(b) = 0$ получаем текущую оценку формантной частоты в виде скользящего среднего:

$$F(t_m) = b = \frac{\sum_{k=1}^{N_{m-1}} F^{(k)}(t_{m-1}) + \sum_{k=1}^{N_m} F^{(k)}(t_m) + \sum_{k=1}^{N_{m+1}} F^{(k)}(t_{m+1})}{N_{m-1} + N_m + N_{m+1}}.$$

Теперь предположим, что частота форманты линейно меняется на каждом отрезке $[t_{m-1}, t_{m+1}]$: $F(t) = at + b$. Тогда оценка форманты, т. е. величина $F(t_m) = at_m + b$, будет зависеть от сетки времен $t_m$. В этом случае функционал МНК имеет вид

$$\Phi(a,b) = \sum_{k=1}^{N_{m-1}} \left(at_{m-1} + b - F^{(k)}(t_{m-1})\right)^2 + \sum_{k=1}^{N_m} \left(at_m + b - F^{(k)}(t_m)\right)^2 + \sum_{k=1}^{N_{m+1}} \left(at_{m+1} + b - F^{(k)}(t_{m+1})\right)^2$$

Условия его минимума $\Phi'_a(a,b) = 0, \Phi'_b(a,b) = 0$ дают систему уравнений для определения параметров *a* и *b*:

$$\begin{cases} A_{11}a + A_{12}b = f_1 \\ A_{21}a + A_{22}b = f_2. \end{cases}$$

Коэффициенты этой системы вычисляются по формулам

$$A_{11} = t_{m-1}^2 N_{m-1} + t_m^2 N_m + t_{m+1}^2 N_{m+1}, A_{12} = A_{21} = t_{m-1} N_{m-1} + t_m N_m + t_{m+1} N_{m+1},$$

$$A_{22} = N_{m-1} + N_m + N_{m+1}; f_1 = t_{m-1} \sum_{k=1}^{N_{m-1}} F^{(k)}(t_{m-1}) + t_m \sum_{k=1}^{N_m} F^{(k)}(t_m) + t_{m+1} \sum_{k=1}^{N_{m+1}} F^{(k)}(t_{m+1}),$$

$$f_2 = \sum_{k=1}^{N_{m-1}} F^{(k)}(t_{m-1}) + \sum_{k=1}^{N_m} F^{(k)}(t_m) + \sum_{k=1}^{N_{m+1}} F^{(k)}(t_{m+1}).$$

Решив эту систему, можно вычислить по найденным числам $a,b$ оценку форманты: $F(t_m) = at_m + b$.

Преимущество такой линейной формулы над скользящим средним состоит в том, что не теряются оценки для первого и последнего импульсов на сегменте с голосовым возбуждением.

В тех случаях, когда формантные частоты меняются мало, оба метода дают очень близкие оценки. В слове /*один*/формантные частоты заметно меняются во времени. Результат применения метода линейной регрессии для первых трех формант в этом случае показан на рисунке 11.

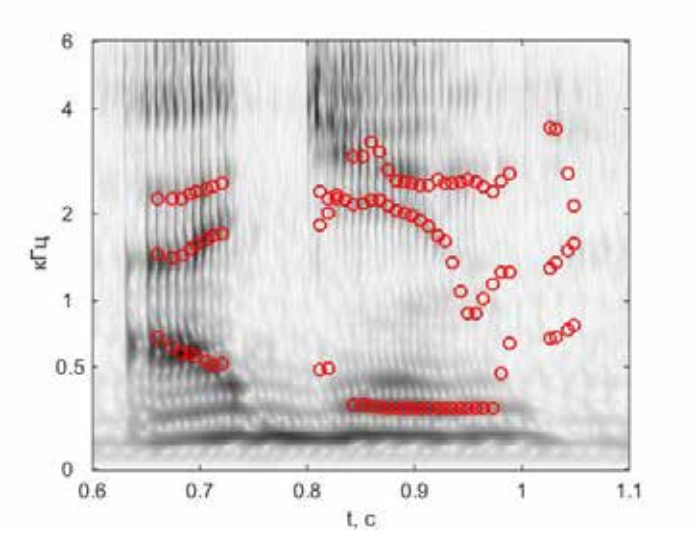


*Рис. 11.* Сонограмма и агрегированные по линейному МНК оценки формантных частот

## 3. Обсуждение

Известны задачи, в которых предъявляются высокие требования к точности оценки периода основного тона $T_0$ и, в частности, его кратковременных модуляций. К ним относятся: диагностика патологии гортани, некоторые неврологические заболевания, распознавание диктора, его эмоционального и физического состояния, дискриминация реального голоса от его синтезированной реплики при идентификации диктора. Существующие методы анализа $T_0$ часто не удовлетворяют этим требованиям. Поэтому необходима разработка новых методов, таких как анализ в фазовой области. Результаты наших исследований, представленных в [11, 12]

и в данной работе, свидетельствуют о значительном разнообразии новых параметров, которые могут использоваться для анализа $T_0$. Также мы установили, что оценки, формируемые этими параметрами, несколько различаются между собой. Из этого следует необходимость создания метода агрегирования оценок $T_0$, полученных в фазовой области, и оценок, традиционно применяемых в речевых технологиях. Проверка эффективности такого метода не должна ограничиваться только синтезированными звуками, поскольку математические модели их сигналов отличаются от волновых процессов в речевом тракте. Для получения объективных характеристик реального речевого сигнала необходимо использовать неинвазивные методы измерения $T_0$, такие, как глоттограммы и ультразвук.

Этот вывод полностью относится и к проблеме оценки моментов начала и конца действия голосового источника или сегментации периода основного тона на интервалы с открытой и закрытой голосовой щелью. Отношение интервала времени с открытой голосовой щелью к текущему периоду основного тона (так называемый фактор Open Quotient) определяется полом диктора, физиологией голосовых складок, физическим и эмоциональным состоянием человека. Погрешность определения момента открытия голосовой щели в фазовой области сильно влияет не только на оценку формантных частот, но и на саму возможность получения такой оценки. Об этом свидетельствуют треки производной по частоте и смешанной производной на рисунках 8 и 10.

Агрегирование коэффициентов автокорреляции фазовых параметров вместе с ограничениями на вариацию соседних периодов основного тона позволяет наиболее надежно определить окрестность границ гласных звуков. Сегментация назальных звуков /*в, л*/и /*р*/ в интервокальной позиции только по агрегированному коэффициенту автокорреляции возможна не всегда. В этих случаях необходима поддержка от анализа в спектрально-временной области.

На сонограмме слова /*шесть*/(Рис. 7) видны форманты в области 2 кГц не только на сегменте гласного, но и на фрикативных /*ш, с* / и даже аспиративного взрыва /*т*/. Треки этих формант заметны также и на фазограммах (Рис. 7б, 9б и 9в). Еще более чётко эти треки проявляются на рисунках 8б и 10б. Более того, на рисунках 7а и 7б видно, что эти треки продолжаются и на смычке /*т*/, что является следствием малой чувствительности фаз к амплитуде. Поэтому оценку формантных частот по фазовым параметрам можно выполнять не только на сегментах с голосовым возбуждением, но и на фрикативных и даже смычках. Это свойство полезно при решении обратных задач восстановления формы речевого тракта, поскольку позволяет оценивать динамику артикуляции с большей уверенностью.

В отличие от агрегированного коэффициента автокорреляции, оценка формантных частот по пикам спектра фазовых параметров не может производиться только в фазовой области. Этому препятствует присутствие паразитных пиков и чувствительность оценок к шумам и искажениям сигнала в акустической среде. Опорные формантные частоты,

найденные независимо от фазового анализа, также могут быть искажены или отсутствовать в какие-то моменты времени. При этом могут появляться и ложные оценки. Поэтому агрегирование фазовых и нефазовых методов оценки формант частот не обязательно должно приводить к непрерывным трекам. В задачах распознавания речи и диктора вполне допустима оценка формантных частот в дискретные, наиболее важные, моменты времени.

Наконец, необходимо высказать замечания относительно технологии сравнения результатов работы алгоритмов анализа речевого сигнала с его разметкой. Популярная в настоящее время разметка фонетистами или даже неспециалистами при машинном обучении нейронных сетей неизбежно будет приводить к серьёзным ошибкам. Во-первых, традиционная фонетическая транскрипция не позволяет описать реально присутствующие в речевом сигнале элементы речевого кода. Во-вторых, такая разметка должна выполняться специально натренированной группой экспертов в рамках достаточно адекватной системы таких элементов.

В [21] приводится описание такой системы, которая содержит около 50 элементов. Оказывается, что в относительно небольшой базе разнообразных сочетаний числительных русского языка для 47 мужчин и женщин наблюдается более 40 вариантов произнесения слова /*шестьдесят*/. Среди них такие сильно различающиеся, как /ШС`ЯТ/, /Ш\ЪСъД`Д`!/, /ШъЗ`Д`Д`!иС`ЯТ/, /ШыЗ`Д`Дh`иС`ЯТТ!h/. Здесь символ ! приписывается символу взрывного и обозначает взрыв в конце смычки, а \/ обозначает так называемый эпентетик, т. е. короткую паузу между глухим фрикативным и последующим звуком с голосовым возбуждением. Символ /ъ/обозначает редуцированный твердый гласный неопределённого фонетического типа. В словосочетании /*шестьдесятдва*/часто происходит озвончение конечного /т/, так что /т/и последующее /д/сливаются, и, соответственно, транскрипция выглядит как /ШыЗ`Д`Д`!иС`ЯДД!ВА/. Такое разнообразие произнесений служит аргументом против обучения нейронных сетей с использованием буквенного представления речевого сигнала и служит предупреждением при оценке погрешности границ сегментов речевого сигнала.

## 4. ЗАКЛЮЧЕНИЕ

Автокорреляционный анализ модуляций производных по времени коэффициента линейной компоненты фазы, максимального значения интервала между нулями разрывной фазовой функции, средней мгновенной частоты в диапазоне частот 1000-3000 Гц и средней энергии амплитудного спектра в этом же диапазоне позволяет оценить период основного тона, моменты начала и конца действия голосового источника и выполнить сегментацию гласных звуков. Формантные частоты гласных определяются путем агрегирования пиков производной фазы по времени, смешанной производной и пиков амплитудного спектра.

### Литература

1. Aarabi P., Shi G., Shanechi M. M., Rabi S. A. Phase-Based Speech Processing. World Scientific Publishing, 2006.
2. Murthy H. A., Yegnanarayna B. Group delay functions and its applications in speech technology // Sadhana. 2011. V. 36. Part 5. P. 745-782.

3. Mowlaee P., Saeidi R., Stylianou Y. Advances in phase-aware signal processing in speech communication // Speech Communication. 2016. V. 81. P. 1-29.
4. 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.
5. Saratxaga I., Sanchez J., Wu Z., Hernaez I., Navas E. Synthetic speech detection using phase information // Speech Communication. 2016. V. 81. P. 30-41.
6. Baghel Sh., Prasanna S. R. M., Guha P. Overlapped speech detection using phase features // JASA. 2021. V. 150. P. 2770-2781
7. Kumar S. B. S, Rao K. S. Voice/non-voice detection using phase of zero frequency filtered speech signal // Speech Communication. 2016. V.81. P. 90-103.
8. Smits R., Yegnanarayana B. Determination of instants of significant excitation in speech using group delay function // IEEE Trans.Speech Audio Process. 1995. V. 3 (5). P. 325-333.
9. 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. N. 2. P. 456-466.
10. Murthy H. A., Yegnanarayna B. Group delay functions and its applications in speech technology // Sadhana. 2011. V. 36. Part 5. P. 745-782.
11. Сорокин В. Н., Леонов А. С. Фазовый анализ активности голосового источника // Акуст. ж. 2021. Т. 67, № 2. С. 185-202.
12. Сорокин В. Н., Леонов А. С. Фазовые модуляции в речевом сигнале // Акуст. ж. 2022. Т. 68, № 2, С. 218-232.
13. Леонов А. С., Сорокин В. Н. Формантный анализ в фазовой области [Электронный ресурс] // Информационные процессы. 2021. Т. 21, № 2. С. 125-134. URL: www.jip.ru.
14. Patterson R. D., Holdsworth J. A functional model of neural activity patterns and auditory images // Advances in Speech, Hearing and Language Processing, 1996. V. 3. P. 547-563.
15. Сорокин В. Н., Цыплихин А. И. Сегментация и распознавание гласных [Электронный ресурс] // Информационные процессы. 2004. Т. 4, № 2. С. 202-220. URL: www.jip.ru.
16. Цыплихин А. И., Сорокин В. Н. Сегментация речи на кардинальные элементы [Электронный ресурс] // Информационные процессы. 2006. Т. 6, № 3. С. 177-207. URL: www.jip.ru.
17. Сорокин В. Н. Детекторы артикуляторных событий // Акуст. ж. 2020. Т. 66, № 1. С. 71-85.
18. Сорокин В. Н. Теория речеобразования. Москва: Радио и Связь, 1985.
19. Сорокин В. Н. Сегментация периода основного тона голосового источника // Акуст. ж. 2016. Т. 62, № 2. С. 247-258.
20. Tsyplikhin A. I. Analysis of Vocal Pulses in a Speech Signal // Acoustical Physics. 2007. V. 53, № 1. P. 105-118.
21. Сорокин В. Н. Речевые процессы. Москва: Народное образование, 2012.
22. Сорокин В. Н. Синтез речи. Москва: Наука, 1992.

## References

1. Aarabi P., Shi G., Shanechi M. M., Rabi S. A. Phase-Based Speech Processing. World Sci-entific Publishing, 2006.
2. Murthy H. A., Yegnanarayna B. Group delay functions and its applications in speech technology // Sadhana. 2011. V. 36. Part 5. P. 745-782.
3. Mowlaee P., Saeidi R., Stylianou Y. Advances in phase-aware signal processing in speech communication // Speech Communication. 2016. V. 81. P. 1-29.

4. Vijayan K., Kumar V., Murty K. S. R. Feature extraction from analytic phase of speech sig-nals for speaker verification // Speaker Odyssey. 2014. P. 1658-1662.
5. Saratxaga I., Sanchez J., Wu Z., Hernaez I., Navas E. Synthetic speech detection using phase information // Speech Communication. 2016. V. 81. P. 30-41.
6. Baghel Sh., Prasanna S. R. M., Guha P. Overlapped speech detection using phase features // JASA. 2021. V. 150. P. 2770-2781
7. Kumar S. B. S, Rao K. S. Voice/non-voice detection using phase of zero frequency filtered speech signal // Speech Communication. 2016. V.81. P. 90-103.
8. Smits R., Yegnanarayana B. Determination of instants of significant excitation in speech using group delay function // IEEE Trans.Speech Audio Process. 1995. V. 3 (5). P. 325-333.
9. 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. N. 2. P. 456-466.
10. Murthy H. A., Yegnanarayna B. Group delay functions and its applications in speech technology // Sadhana. 2011. V. 36. Part 5. P. 745-782.
11. Sorokin V. N., Leonov A. S. Fazovyj analiz aktivnosti golosovogo istochnika // Akust. zh. 2021. T. 67, № 2. S. 185-202.
12. Sorokin V. N., Leonov A. S. Fazovye modulyacii v rechevom signale // Akust. zh. 2022. T. 68, № 2, S. 218-232.
13. Leonov A. S., Sorokin V. N. Formantnyj analiz v fazovoj oblasti [Elektronnyj resurs] // Informacionnye processy. 2021. T. 21, № 2. S. 125-134. URL: www.jip.ru.
14. Patterson R. D., Holdsworth J. A functional model of neural activity patterns and auditory images // Advances in Speech, Hearing and Language Processing, 1996. V. 3. P. 547-563.
15. Sorokin V. N., Cyplihin A. I. Segmentaciya i raspoznavanie glasnyh [Elektronnyj resurs] // Informacionnye processy. 2004. T. 4, № 2. S. 202-220. URL: www.jip.ru.
16. Cyplihin A. I., Sorokin V. N. Segmentaciya rechi na kardinal'nye elementy [Elek-tronnyj resurs] // Informacionnye processy. 2006. T. 6, № 3. S. 177-207. URL: www.jip.ru.
17. Sorokin V. N. Detektory artikulyatornyh sobytij // Akust. zh. 2020. T. 66, № 1. S. 71-85.
18. Sorokin V. N. Teoriya recheobrazovaniya. Moskva: Radio i Svyaz', 1985.
19. Sorokin V. N. Segmentaciya perioda osnovnogo tona golosovogo istochnika // Akust. zh. 2016. T. 62, № 2. S. 247-258.
20. Tsyplikhin A. I. Analysis of Vocal Pulses in a Speech Signal // Acoustical Physics. 2007. V. 53, № 1. P. 105-118.
21. Sorokin V. N. Rechevye processy. Moskva: Narodnoe obrazovanie, 2012.
22. Sorokin V. N. Sintez rechi. Moskva: Nauka, 1992.

## ANALYSIS OF VOWELS IN THE PHASE DOMAIN

***Sorokin V. N.,*** *Doctor of Physical and Mathematical Sciences Institute for Information Transmission Problems, Russian Academy of Sciences, principal scientist Moscow, Russia,*
*vns@iitp.ru*

***Leonov A. S.,*** *Doctor of Physical and Mathematical Sciences, Professor, National Research Nuclear University «MEPHI» Moscow, Russia*

Algorithms and numerical experiments for determining the parameters of a speech signal in the phase domain are presented. A method for aggregating the results of autocorrelation analysis of modulations of the time derivatives of the coefficient of the

linear phase component, the maximum value of the interval between zeros of the discontinuous phase function, the average instantaneous frequency in the frequency range of 1000-3000 Hz, and the average energy of the amplitude spectrum in the same range is described. The period of the fundamental tone, the moments of the beginning and end of the vocal source activity are estimated, and segmentation of vowel-like sounds is also performed. The formant frequencies of vowels are determined by aggregating the peaks of the phase derivative with respect to time, the mixed derivative, and the peaks of the amplitude spectrum.

*• Speech analysis • group delay • instantaneous frequency • fundamental tone parameters • formant frequencies*