АКУСТИЧЕСКИЙ ЖУРНАЛ, 2016, том 62, № 2, с. 216–226

ОБРАБОТКА АКУСТИЧЕСКИХ СИГНАЛОВ. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

УДК 004.934:519.614:004.42

МОДЕЛЬ ПРОЦЕССА СИНГУЛЯРНОГО ОЦЕНИВАНИЯ ЧАСТОТЫ ОСНОВНОГО ТОНА РЕЧЕВОГО СИГНАЛА

© 2016 г. Д. А. Вольф, Р. В. Мещеряков

Томский государственный университет систем управления и радиоэлектроники
634050 Томск, пр. Ленина 40
E-mail: runsolar@mail.ru

Поступила в редакцию 18.02.2015 г.

Предложен новый способ оценивания частоты основного тона речи, основанный на сингулярном спектральном анализе. Представлена сингулярная модель вокализированного сегмента речевого сигнала, в которой рассматривается прямая и обратная задачи. Проведено исследование процесса сингулярного оценивания частоты основного тона речи. Выполнены экспериментальные исследования с моделью, где даны оценки адекватности и достоверности. Введено понятие сингулярного оценивания частоты основного тона речи.

Ключевые слова: сингулярный спектральный анализ речи, частота основного тона речи, модель, численная реализация, оценка адекватности и достоверности.

DOI: 10.7868/S0320791916020143

ВВЕДЕНИЕ

Анализ речевого сигнала в настоящее время является одной из перспективных областей исследований в области акустики. Предметом данной работы является описание модели процесса сингулярного оценивания одного из основных параметров устной речи: частоты колебаний голосовых складок, называемой основным тоном – *F*0. В настоящее время популярными алгоритмами оценивания частоты основного тона (ЧОТ) речи являются SHS[1], RAPT [2], YIN [3] и SWIPE' [4]. Популярность перечисленных алгоритмов обусловлена хорошей функциональностью, низким процентом грубых ошибок и наличием свободно распространяемых версий их реализаций. Большинство современных измерителей основного тона состоят из трех основных модулей (рис. 1) [5]: 1 – модуль предобработки или приведения сигнала к требуемым характеристикам, 2 – генератор кандидатов действительного искомого периода основного тона, 3 – модуль постобработки или выбора наилучшего кандидата с последующим уточнением значения частоты основного тона.

Главным недостатком данного класса алгоритмов является их зависимость от точности нахождения пиков. Наличие пиков и их амплитуда зависят от длины и вида окна анализа, а также от типа звука, что довольно часто приводит к ошибкам. Более того, точность зависит от значения частоты основного тона и от частоты дискретизации [6]. Еще одно ограничение обусловлено периодической моделью сигнала, лежащей в их основе, которая подразумевает точное повторение периода основного тона и не допускает его изменения на протяжении окна анализа. Например, при появлении модуляций – изменений частоты основного тона, точность оценок также существенно снижается. При исследованиях речевых сигналов обычно используется математический аппарат спектрального анализа Фурье или вейвлет-анализ. В настоящей работе применен аппарат сингулярного спектрального анализа (ССА "Гусеница"), разработанного и обоснованного в конце XX века сотрудниками С.-Петербургского государственного университета [7, 8]. В зарубежной литературе описан широкий класс методов, алгоритмически и идейно близких к методу "Гусеница", метод известен как Singular Spectrum Analysis (SSA) [9–11]. Он основан на анализе главных компонент и позволяет исследовать стационарные и нестационарные временные ряды. Связь между классическими методами анализа стационарных временных рядов и методом главных компонент рассмотрена в работах Бриллинджера [12]. Например, в работе Bagshaw [13] утверждается, что методы, работающие во временной области, обладают наименьшей ошибкой по сравнению с другими частотными методами принятия решения о присутствии голоса в речи – не более 17%. В [14] показано, что такие методы являются робастными в отношении принятия решения о вокализованном или невокализованном сегменте речи в условиях шума, искажений и побочных помех в сигнале.

Целью данной работы является получение новой технологии оценивания ЧОТ речи с учетом амплитудно-частотных модуляций. Ее новизна

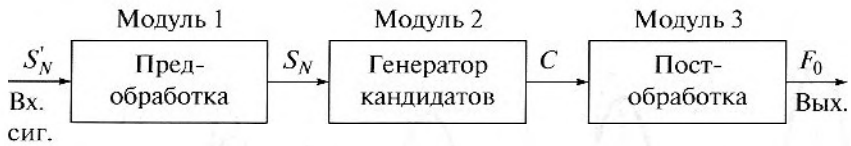

Рис. 1. Схема типового оценщика основного тона.

заключается в применении математического аппарата сингулярного спектрального анализа к речевым сигналам.

СИНГУЛЯРНЫЙ АНАЛИЗ РЕЧЕВОГО СИГНАЛА

Несмотря на то, что речеобразующий механизм представляет собой относительно труднодоступную систему, в некотором приближении рассмотрим образование вокализированных сегментов речи. Одним из источников образования звуков является голосовой источник, который возникает при колебании голосовых складок. Он участвует в образовании нескольких групп звуков и по степени участия голосового источника звуки делятся на гласные и согласные. Для вокализированного сегмента речи экспериментально было установлено, что на "фонетический смысл" гласных звуков существенно влияют амплитуды на ЧОТ и обертоновых составляющих речевого сигнала. Однако даже на современном этапе представляется весьма сложным получить точные данные всех параметров речевого тракта.

Рассмотрим модель вокализованного сегмента речевого сигнала применительно к задачам анализа и синтеза речи. Входной сигнал $x(t)$ поступает от голосовых складок (природный квазигармонический генератор – генеративная система), проходит через N параллельно соединенных резонаторов, характеризующих форму речевого тракта, и на выходе формируется определенный произносимый вокализованный речевой сегмент $y(t)$. Таким образом, математическую модель вокализованного речевого сегмента можно описать в виде суммы некоторого набора амплитудных, фазовых и частотных параметров, формируемых в результате прохождения полигармонического колебания через резонансную систему [6]

$$S(t) = \sum_{n=0}^{N-1} I_n(t) \sin\left((n+1) \int_0^t \omega_0(\tau) d\tau + \varphi_n \right),$$

где $n = 0, 1, 2, \ldots$ – номер гармоники основного тона; I_n – амплитуды гармоник; ω_0 – частота основного тона (рад/с); φ_n – начальная фаза гармоник; $S(t)$ – конечный продукт генеративной и резонансной системы.

Для выделения генеративной и резонансной составляющей представляет интерес конечный продукт $S(t)$ [15]. Решение задачи может быть основано на анализе импульсных характеристик резонансной системы для распознавания или дальнейшего синтезирования речи диктора и т.д. Модели речеобразования и модели речевых сигналов рассмотрены Сорокиным В.Н. в [16].

Рассмотрим прямую задачу. Пусть временной ряд S_N – ряд, полученный в результате процедуры дискретизации речевого сигнала $S(t)$. Ряд S_N назовем фонемным. Осуществим с фонемным рядом процедуру ганкелизации [8]

$$A = [S_{i-1}, \ldots, S_{i+L-1}]^{\mathrm{T}}, \quad 1 \le i \le K, \quad K = N - L + 1, \quad (1)$$

и получим траекторную матрицу A, состоящую из K векторов вложения длины L. Траекторную матрицу (1) можно представить в виде матричного разложения

$$A = \sum_{i=0}^{L-1} A^{\langle i \rangle} = \sum_{i=0}^{L-1} (\sqrt{\lambda_i}\mathbf{u}^{\langle i \rangle})[\mathbf{x}^{\langle i \rangle}]^{\mathrm{T}}, \quad (2)$$

где λ_i – i-е собственное значение ковариационной матрицы AA^{T}; $\mathbf{u}^{\langle i \rangle}$ – i-й собственный вектор ковариационной матрицы AA^{T}; $\mathbf{x}^{\langle i \rangle}$ – i-й собственный вектор, образованный строками матрицы A. Для произведения векторов в (2) матричное усреднение по диагонали равно

$$T_j^{\langle n \rangle} = \begin{cases} \dfrac{1}{j+1} \displaystyle\sum_{i=0}^{j} [\sqrt{\lambda}\mathbf{u}\mathbf{x}^{\mathrm{T}}]_{iK+j-i}^{\langle n \rangle}, & 0 \le j < L, \\ \dfrac{1}{L} \displaystyle\sum_{i=0}^{L-1} [\sqrt{\lambda}\mathbf{u}\mathbf{x}^{\mathrm{T}}]_{iK+j-i}^{\langle n \rangle}, & L \le j < K, \\ \dfrac{1}{N-j} \displaystyle\sum_{i=0}^{L-1-(j-K)} [\sqrt{\lambda}\mathbf{u}\mathbf{x}^{\mathrm{T}}]_{(j-K+i)K+K-1-i}^{\langle n \rangle}, & K \le j < N, \end{cases} \quad (3)$$

что позволяет описать двухмерный массив данных $T_{L,N}$, в строках которого содержится квазигармонический спектр (рис. 2).

Аналогично тому, как в гармонических моделях осуществляется проекция в гармоническом базисе (например, в преобразованиях Фурье), так

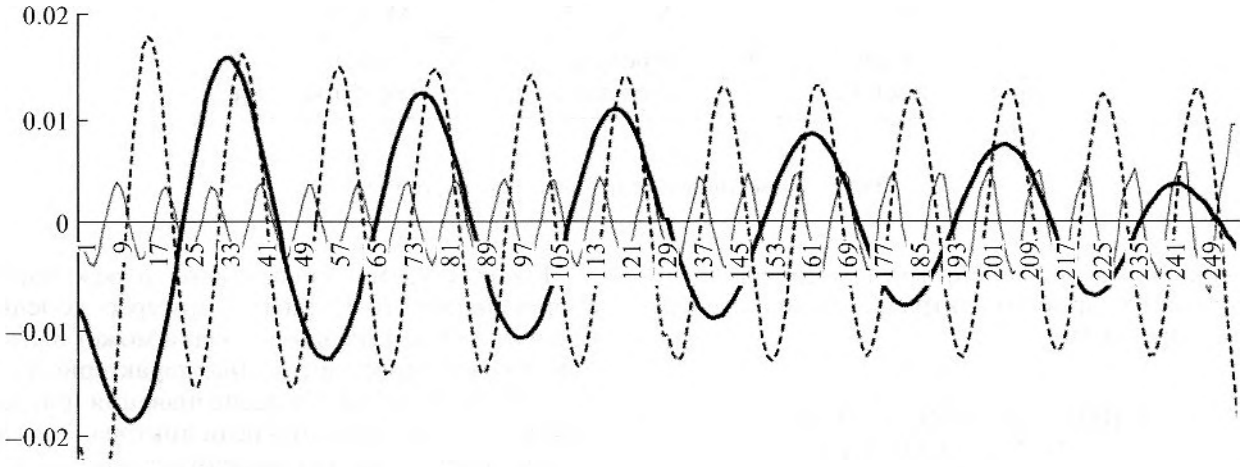

Рис. 2. Разложение фонемы "е" в первые три квазигармонические компоненты (субфонемы) на основе сингулярного спектрального анализа.

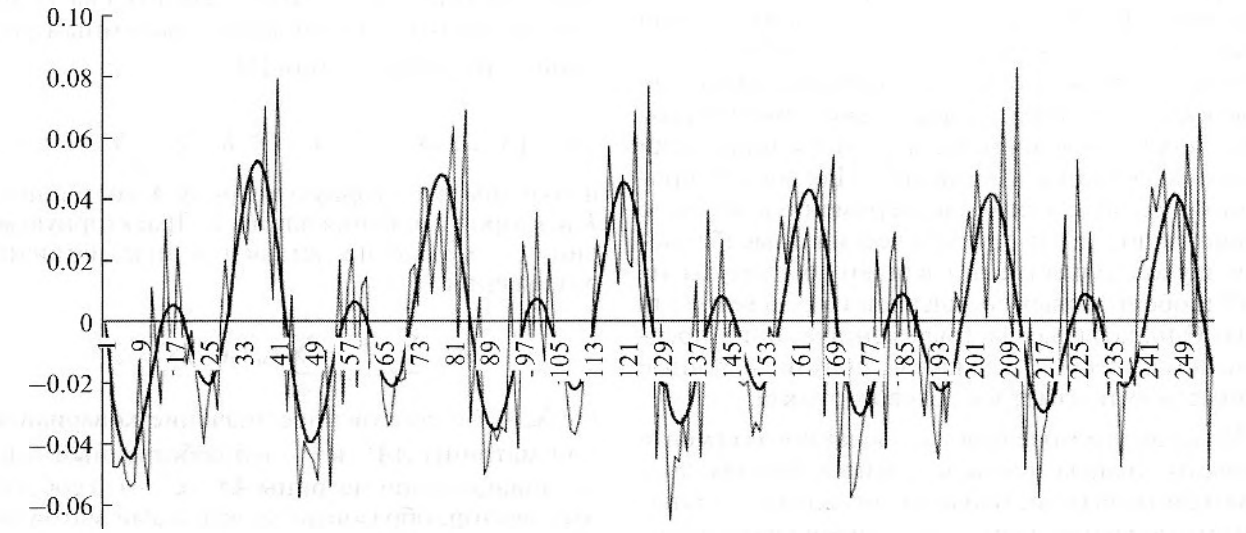

Рис. 3. Реконструкция исходного речевого ряда S_N звука "е" по первым трем квазигармоникам.

и в прямой задаче сингулярного спектрального анализа речи осуществляется проекция в базисе собственных векторов, порождаемых столбцами ковариационной матрицы AA^{T} и строками матрицы A. При этом собственные векторы $\mathbf{u}^{\langle i\rangle}$, $\mathbf{x}^{\langle i\rangle}$ имеют квазигармоническую структуру.

Рассмотрим обратную задачу. Сумма соответствующих j-х квазигармоник ряда (3) равна исходному фонемному ряду (рис. 3):

$$S_N = \sum_{n=0}^{L-1} T_j^{\langle n\rangle}, \quad j = 0, \ldots, N-1. \tag{4}$$

Пусть для некоторой последовательности $i = 0, 1, \ldots$ собственные числа λ_i, $\mathbf{u}^{\langle i\rangle}$, $\mathbf{x}^{\langle i\rangle}$ являются эмпирически найденными величинами, образующими совокупность параметров для образования звуков речи. Тогда для произведения

$$A_i = \sqrt{\lambda_i}\mathbf{u}^{\langle i\rangle}[\mathbf{x}^{\langle i\rangle}]^{\mathrm{T}}, \quad i = 0, \ldots \tag{5}$$

выражение (3) можно принять в качестве синтезатора акустических сигналов, генерируемых речеобразующим трактом. Очевидно, что произведение (5) является входным параметром для (3). Полученную матрицу (5) будем считать неявной квазитраекторной, так как элементы, расположенные на антидиагоналях, представляют собой разброс возле некоторой средней величины.

Синтезирование величин λ_i, $\mathbf{u}^{\langle i\rangle}$, $\mathbf{x}^{\langle i\rangle}$ в качестве резонаторов речеобразующего тракта без решения прямой задачи является достаточно сложной задачей. Тем не менее, для построения модели сингулярного оценивания ЧОТ речи систему

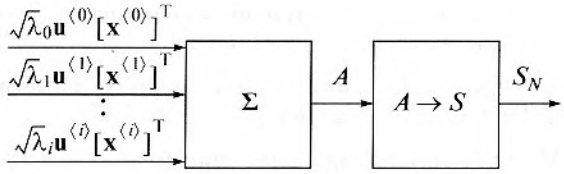

Рис. 4. Схема модели сингулярного синтезатора речи.

$$\begin{cases} T_j^{\langle n\rangle} = \\ \quad = \begin{cases} \dfrac{1}{j+1}\displaystyle\sum_{i=0}^{j}[\sqrt{\lambda}\mathbf{u}\mathbf{x}^{\mathrm{T}}]_{iK+j-i}^{\langle n\rangle}, & 0 \le j < L, \\ \dfrac{1}{L}\displaystyle\sum_{i=0}^{L-1}[\sqrt{\lambda}\mathbf{u}\mathbf{x}^{\mathrm{T}}]_{iK+j-i}^{\langle n\rangle}, & L \le j < K, \\ \dfrac{1}{N-j}\displaystyle\sum_{i=0}^{L-1-(j-K)}[\sqrt{\lambda}\mathbf{u}\mathbf{x}^{\mathrm{T}}]_{(j-K+i)K+K-1-i}^{\langle n\rangle}, & K \le j < N, \end{cases} \\ S_N = \displaystyle\sum_{n=0}^{L-1} T_j^{\langle n\rangle}, \quad j = 0,\ldots,N-1, \end{cases} \tag{6}$$

примем в качестве сингулярной модели вокализированного сегмента речевого сигнала (рис. 4).

Таким образом, для определения частоты основного тона речи формулируются две задачи: 1) разложение исходного речевого сигнала в сингулярный спектр квазигармонических компонент; 2) выбор квазигармонической составляющей соответствующей частоте основного тона речи.

МОДЕЛЬ ПРОЦЕССА СИНГУЛЯРНОГО ОЦЕНИВАНИЯ ЧОТ

Представим решение задачи в параметрическом виде

$$(T0_N, F0, Amp) = \mathrm{SEPT}(S_N),$$

что позволит из сингулярной модели вокализированного сегмента речевого сигнала (6), сформулировать эвристическое описание модели сингулярного оценивания ЧОТ речи. Акустический сигнал в виде дискретного фонемного ряда S_N поступает на вход генератора сингулярного спектра (ГСС). На выходе ГСС формируется квазигармонический спектр в виде двухмерного массива данных $T_{L,N}$. Квазигармонический спектр $T_{L,N}$ поступает на вход системы выбора спектральной составляющей, соответствующей частоте основного тона речи. Данные подаются на первый вход блока управления матрицей временного спектра (УМВС). Далее квазигармонический спектр $T_{L,N}$ с первого выхода УМВС поступает в блок измерения частоты временного спектра (ИЧВС). В блоке ИЧВС с использованием измерения обратной величины среднего периода по максимумам (в иных случаях подсчитывается число переходов через нуль) решается задача измерения частот квазигармонического спектра. Параллельно в ИЧВС осуществля-

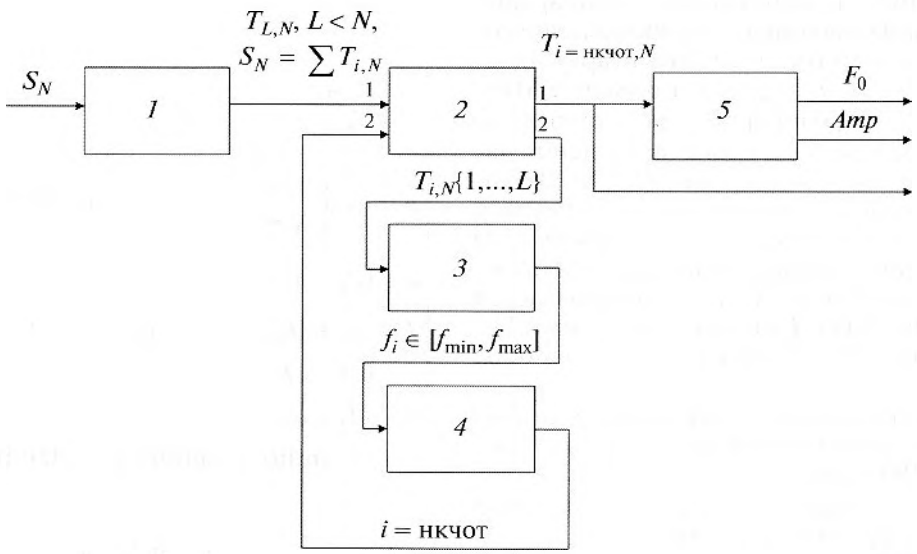

Рис. 5. Базовая схема модели процесса для сингулярного измерения мгновенной частоты основного тона речевого сигнала. Состав блоков: *1* – блок генератора сингулярного спектра (ГСС); *2* – блок управления матрицей временного спектра (УМВС); *3* – блок измерения частоты временного спектра (ИЧВС); *4* – блок выбора номера компоненты с частотой основного тона речи (ВНКЧОТ); *5* – блок вычисления частоты и амплитуды основного тона речи (ВЧА). Состав сигналов: N – длина анализируемого кадра; L – количество спектральных компонент; S_N – дискретный речевой сигнал; $T_{L,N}$ – временной спектр (сингулярный спектр речевого сигнала); f_i – частота; i – номер компоненты с частотой основного тона речи; $T_{i=\text{нкчот},N}$ – компонента с частотой основного тона речи; $F0$ – частота основного тона речи; Amp – амплитуда.

ется уменьшение плотности (размерности) квазигармонического спектра $T_{L,N}$ с L до величины $M < L$, т.е. ряд $T_{L,N}$ разрежается до размерности M. Это обеспечивает сужение границ поиска квазигармонического сигнала, соответствующего ЧОТ

$$f_0 \in [f_{\min}, f_{\max}],$$

где f_0 – ЧОТ речи; $f_{\min}, f_{\max}$ – соответственно минимальная и максимальная граница интервала существования ЧОТ.

С выхода блока ИЧВС совокупность частот (частотный ряд)

$$f_{\min} \le \{f_1, f_3, f_j, \ldots, f_M\} \le f_{\max}$$

поступает на вход блока выбора номера компоненты ЧОТ речи (ВНКЧОТ), причем индекс множества сохраняется прежним. По данному индексу осуществляется выбор компоненты ЧОТ речи из массива данных $T_{L,N}$, т.е.

если $T_{i,N}\{i = 1, 2, \ldots, L;\ i \in I\}$, а $f_{\min} \le \{f_1, f_3, f_j, \ldots, f_M\} \le f_{\max}$, то $J \in I$.

Из полученного ряда частот $\{f_1, f_3, f_j, .., f_M\}$ в блоке ВНКЧОТ осуществляется выбор ЧОТ речи и определяется соответствующий ему индекс j из множества $J \in I$, равный номеру компоненты с частотой основного тона (НКЧОТ). Выбор ЧОТ речи в блоке ВНКЧОТ определяется критерием наименьшей кратной частотной величины основного тона:

$$f_0 \in \{\min(f_j),\ 2\min(f_j), \ldots, M\min(f_j)\},$$

причем $f_j \in (f_{\min} \le \{f_1, f_2, \ldots, f_M\} \le f_{\max})$.

С выхода блока ВНКЧОТ индекс, отвечающий номеру компоненты ЧОТ, поступает на второй вход блока УМВС, где выбираются (активируются) строки матрицы (таблицы) со спектральной составляющей, соответствующей квазигармоническому сигналу с ЧОТ $T_{i=\text{нкчот},N}$. На выходе УМВС сигнал с ЧОТ $T_{i=\text{нкчот},N}$ одновременно поступает на выход из системы и на вход блока вычисления частоты и амплитуды основного тона (ВЧА). В блоке ВЧА вычисляются максимальные значения ряда $T_{i=\text{нкчот},N}$ и подсчитывается количество $(m-1)$ обратных величин, равных периодам, расположенным в данном ряде. Далее подсчитывается средняя величина ЧОТ $F0$ и амплитуда Amp. Результаты значений $F0$ и Amp поступают на выходы системы.

Исходя из эвристического описания, рассмотрим базовую схему модели процесса сингулярного оценивания ЧОТ (рис. 5):

1 – блок ГСС, выполняющий сингулярный спектральный анализ речевого сигнала;

2 – блок УМВС хранения результатов сингулярного анализа речевого сигнала в виде спектральных составляющих;

3 – блок ИЧВС, реализующий отбор спектральных составляющих из интервала частот, в котором расположена ЧОТ речи;

4 – блок ВНКЧОТ речи, в котором по алгоритму наименьшей кратной величины частоты основного тона проводится выбор номера компоненты, соответствующей ЧОТ;

5 – блок ВЧА вычисления средней величины ЧОТ $F0$ и амплитуды Amp.

Модели генератора сингулярного спектра речевого сигнала и выбора квазигармонической составляющей ЧОТ речи описываются системами вида:

$$\begin{cases} A = \begin{pmatrix} S_0 & S_1 & S_2 & \ldots & S_{K-1} \\ S_1 & S_2 & S_3 & \ldots & S_K \\ S_2 & S_3 & S_4 & \ldots & S_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_{L-1} & S_L & S_{L+1} & \ldots & S_{N-1} \end{pmatrix}; \\ C = A^{\mathrm{T}}A; \\ \mathbf{V}_A^{\mathrm{T}} = \mathbf{U}_C^{\mathrm{T}} A D_C^{-1}; \\ T_j^{\langle n \rangle} = \\ = \begin{cases} \dfrac{1}{j+1}\displaystyle\sum_{i=0}^{j} [\sqrt{\lambda}\mathbf{u}\mathbf{v}^{\mathrm{T}}]_{iK+j-i}^{\langle n \rangle}, & 0 \le j < L, \\ \dfrac{1}{L}\displaystyle\sum_{i=0}^{L-1} [\sqrt{\lambda}\mathbf{u}\mathbf{v}^{\mathrm{T}}]_{iK+j-i}^{\langle n \rangle}, & L \le j < K, \\ \dfrac{1}{N-j}\displaystyle\sum_{i=0}^{L-1-(j-K)} [\sqrt{\lambda}\mathbf{u}\mathbf{v}^{\mathrm{T}}]_{(j-K+i)K+K-1-i}^{\langle n \rangle}, & K \le j < N. \end{cases} \end{cases} \quad (7)$$

$$\begin{cases} f_n = \dfrac{p}{N\Delta t}, \\ p = \left\{ k, \left\| \left[\dfrac{1}{N}\displaystyle\sum_{j=0}^{N-1} T_j^{\langle n \rangle} e^{\frac{-2\pi i}{N}kj} \right]_k \right\| \subseteq MAX, k = \overline{0, N-1} \right\}, \\ n = \overline{0, L-1}, \\ f_j = f_n \in [f_{\min} \le f_n \le f_{\max}], n = \overline{0, L-1}, \\ j = 0, 1, \ldots, K < L, \\ f_0 = f_{j=\text{нкчот}} = \\ = f_j \in \{\min(f_j),\ 2\min(f_j), \ldots, M\min(f_j)\}, \\ j = \overline{1, K}, \\ T0_n = T_{j=\text{нкчот},n}, \quad n = \overline{0, N-1}, \\ F0 = \dfrac{1}{m-1}\displaystyle\sum_{i=1}^{m} \frac{1}{(k_{i-1} - k_i)\Delta t}, \\ k_i = \{n, T0_n \subset MAX, n = \overline{0, N-1}\}, \quad i = \overline{1, m}, \\ Amp = \dfrac{1}{m}\displaystyle\sum \max(T0_n), \quad n = 1, 2, \ldots, m. \end{cases} \quad (8)$$

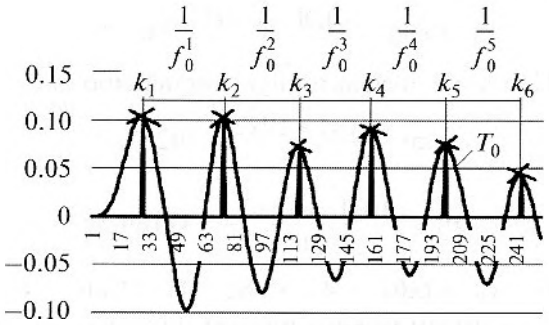

Рис. 6. Визуализация параметров ряда $T0$, по которым вычисляется средняя амплитуда и ЧОТ.

Здесь S_N – исходный временной ряд; N – длина ряда; L – размер спектрального окна; A – траекторная матрица наблюдений; C – бисимметричная матрица; $\mathbf{U}_C$ – левая сингулярная матрица поворота, определяемая выражением

$$\mathbf{U}_C = \begin{pmatrix} u_0^0 & u_0^1 & u_0^2 & \dots & u_0^{L-1} \\ u_1^0 & u_1^1 & u_1^2 & \dots & u_1^{L-1} \\ u_2^0 & u_2^1 & u_2^2 & \dots & u_2^{L-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{L-1}^0 & u_{L-1}^1 & u_{L-1}^2 & \dots & u_{L-1}^{L-1} \end{pmatrix};$$

$\mathbf{V}_A^{\mathrm{T}}$ – правая сингулярная матрица поворота; $\mathbf{u}^{\langle n\rangle}$ – левый сингулярный вектор; $\mathbf{v}^{\langle n\rangle}$ – правый сингулярный вектор; D – диагональная матрица, состоящая из собственных значений λ_i бисимметричной матрицы C и края спектра значений исходной матрицы A

$$D_C = \mathrm{diag}\{\lambda_0, \lambda_1, \lambda_2, \dots, \lambda_{L-1}\},\ \lambda_0 < \lambda_1 <, \dots, < \lambda_{L-1};$$

T_i^n – спектр временных рядов; f_n – одномерное частотное представление временного спектра T_i^n при условии, что искомая ЧОТ f_0 содержится в интервале

$$f_0 \in \{\min(f_i), 2\min(f_i), \dots, M\min(f_i)\};$$

p – индекс элемента в ряде T_i^n, соответствующий максимальной амплитуде в n-й квазигармонике; Δt – величина, обратная частоте дискретизации; $T0_N$ – временной ряд, соответствующий квазигармонике с ЧОТ речи; $F0$ – средняя частота основного тона речи

$$F0 = \frac{f_0^1 + f_0^2 + \dots + f_0^m}{m-1},$$

где $(m-1)$ – количество обратных величин, равных периодам, расположенным в ряде $T0_N$ (f_0^i – локальная частота тона);

$$f_0^1 + f_0^2 + \dots + f_0^m = \frac{1}{(k_2 - k_1)\Delta t} + \frac{1}{(k_3 - k_2)\Delta t} + \dots + \frac{1}{(k_m - k_{m-1})\Delta t} = \sum_{i=1}^{m} \frac{1}{(k_i - k_{i-1})\Delta t},$$

где k_i – номер индекса в точке максимума (рис. 6)

$$k_i = \{n, T0_n \subset \max, n = \overline{0, N-1}\},\quad i = \overline{1, m};$$

Amp – средняя амплитуда гармоники (средняя величина по максимумам в ряде $T0_N$), соответствующая ЧОТ речи.

Система (7) решает задачу разложения исходного речевого сигнала (одномерного ряда) S_N в спектр квазигармонических компонент (двумерная матрица) T_i^n, $[i = 0, 1, \dots, N-1; n = 0, 1, \dots, L-1]$.

Система (8) решает задачу выбора квазигармонической составляющей соответствующей ЧОТ речи:

– вычисление одномерного ряда $T0_N$, соответствующего ЧОТ речи, из многомерного ряда T_i^n, соответствующего временному спектру;

– оценивание средней ЧОТ речи $F0$ (с учетом модуляции), содержащейся во временном ряде $T0_N$;

– определение средней амплитуды во временном ряду $T0_N$.

ОЦЕНКА АДЕКВАТНОСТИ И ДОСТОВЕРНОСТИ МОДЕЛИ

Оценим эффективность модели сингулярного оценивания ЧОТ. Для этого проведем вычислительный эксперимент (дисперсионный анализ) [17] согласно следующей схеме:

1) Пусть имеются две независимые выборки от одного источника данных. Считаем, что данные в этих выборках приближены к нормальному распределению. В качестве нулевой гипотезы H_0 примем положение, что различия между оценками ЧОТ, полученными с помощью программной реализации модели сингулярного оценивания "Septv1" и программы "Praat" [18, 19], несущественны, т.е. различен лишь способ измерения, а результаты измерения имеют несущественные различия.

2) В качестве исходных данных выберем фонемные ряды гласных звуков английской речи: $\{[\mathrm{a}]_i\}^{256}$, $\{[\mathrm{e}]_i\}^{256}$, $\{[\mathrm{ɵ}]_i\}^{256}$, $\{[\mathrm{i}]_i\}^{256}$, $\{[\mathrm{o}]_i\}^{256}$, $\{[\mathrm{u}]_i\}^{256}$, $\{[\mathrm{ɨ}]_i\}^{256}$, $\{[\mathrm{ɪ}]_i\}^{256}$, $\{[\mathrm{ʉ}]_i\}^{256}$, $\{[\mathrm{æ}]_i\}^{256}$;

3) С помощью программы "Septv1" (выборка $x1_i$) и программы "Praat" (выборка $x2_i$) проведем оценки ЧОТ (таблица 2);

Таблица 1. Независимые оценки (выборки) ЧОТ речи

Звук	i	$x1_i$, Гц	$x2_i$, Гц
		1	2
[a]	1	203.13	203.4
[e]	2	193.96	194
[ɵ]	3	199.80	200.7
[i]	4	203.5	204
[o]	5	213.36	212.5
[u]	6	214.27	212.3
[ɨ]	7	204.86	207
[I]	8	201.05	202.2
[ʉ]	9	204.80	206.5
[æ]	10	187.31	187.2
$\overline{x1}$, $\overline{x2}$		202.60	202.98
$\overline{x1}^2$, $\overline{x2}^2$		41106.54	41255.15

4) На основании табличных данных (табл. 1) вычислим эмпирическое корреляционное соотношение;

5) На основании полученной эмпирической оценки примем или отклоним нулевую гипотезу H_0.

Используя исходные данные из табл. 1, рассчитаем средние значения и средние квадраты для векторов (факторов 1, 2):

$$x1 = [203.13,\ 193.96,\ 199.80,\ 203.5,\ 213.36,\ 214.27,\ 204.86,\ 201.05,\ 204.80,\ 187.31];$$

$$x2 = [203.4,\ 194.0,\ 200.7,\ 204.0,\ 212.5,\ 212.3,\ 207.0,\ 202.2,\ 206.5,\ 187.2],$$

$$\overline{x1} = \frac{1}{10}\sum_{i=1}^{10} x1_i = 202.60,$$

$$\overline{x2} = \frac{1}{10}\sum_{i=1}^{10} x2_i = 202.98,$$

$$\overline{x1}^2 = \frac{1}{10}\sum_{i=1}^{10} x1_i^2 = 41106.54,$$

$$\overline{x2}^2 = \frac{1}{10}\sum_{i=1}^{10} x2_i^2 = 41255.15.$$

Вычислим факторную дисперсию:

$$D1 = \overline{x1}^2 - (\overline{x1})^2 = 41106.54 - 41048.38 = 58.17,$$

$$D2 = \overline{x2}^2 - (\overline{x2})^2 = 41255.15 - 41200.88 = 54.27.$$

Определим среднюю внутрифакторную дисперсию:

$$D\text{свд} = \frac{10D1 + 10D2}{20} = 56.22.$$

Найдем общефакторную дисперсию:

$$\overline{x0} = \frac{10\overline{x1} + 10\overline{x2}}{20} = 202.79,$$

$$\overline{x0}^2 = \frac{10\overline{x1}^2 + 10\overline{x2}^2}{20} = 41180.85,$$

$$D0 = \overline{x0}^2 - (\overline{x0})^2 = 41180.85 - 41124.60 = 56.25.$$

Рассчитаем межфакторную дисперсию по формуле суммы разности общефакторной и факторной дисперсий:

$$D\text{мф} = \frac{10(D0 - D1) + 10(D0 - D2)}{20} = 0.035.$$

Вычислим эмпирическое корреляционное соотношение:

$$\eta = \sqrt{\frac{D\text{мф}}{D0}} = \sqrt{\frac{0.035}{56.25}} = 0.025.$$

Относительно шкалы Чеддока [20] разница между выборками $x1_i$ и $x2_i$ слабая, всего 2.5%, следовательно, нет оснований отвергать нулевую гипотезу H_0. Таким образом, принимается гипотеза о незначительных различиях между оценками ЧОТ, полученными с помощью программ "Septv1" и "|Praat". Для 100 несортированных (как мужских, так и женских) образцов вокализированных сегментов речи из базы данных Disordered Voice Database [21] дисперсионный анализ показал идентичные результаты.

При проведении вычислительного эксперимента появился фактор, который требует дополнительного изучения. Необходимо учитывать не только канал анализа речевого сигнала, но и канал синтеза [22]. В статье [23] описывается постановка эксперимента по оценке параметров голосового источника. В результате эксперимента рассматривается распределение периодов основного тона женских и мужских голосов на ударных гласных числительных русского языка и их аппроксимация гамма-распределением.

Если принять, что множества частотных выборок основных тонов $i = 1, ..., N: xn_i$ для женских и $i = 1, ..., N: ym_i$ для мужских дикторов, при [$n = 1, ...,$ б.ч.; $m = 1, ...,$ б.ч.] (б.ч. – большое число), имеют некоторую сходимость к нормальному гамма-распределению, то можно предположить, что нормальный диапазон ЧОТ для любого диктора (или же диапазон, характеризующий конкретного диктора) составляет

$$[\overline{xn} - 2\sqrt{D}, \overline{xn} + 2\sqrt{D}],$$

$$[\overline{ym} - 2\sqrt{D}, \overline{ym} + 2\sqrt{D}],$$

где: $\overline{xn}$, $\overline{ym}$ – средняя величина ЧОТ для женского и мужского диктора соответственно по всему диапазону гласных звуков речи; n, m – порядковый номер диктора; D – дисперсия (в квадратных ча-

Таблица 2. Оценка ЧОТ с использованием синтетических сигналов

		Скорость изменения ЧОТ, Гц/мс				
		0	0.5	1.0	1.5	2.0
HNR 25 дБ						
RAPT	GPE	0.000	0.000	0.000	3.10	8.31
	MFPE	0.052	0.189	0.523	1.245	2.208
YIN	GPE	0.000	0.000	0.000	0.000	1.38
	MFPE	0.041	0.173	0.452	0.802	1.219
SWIPE'	GPE	0.000	0.000	0.000	0.000	0.000
	MFPE	0.035	0.14	0.289	0.413	0.712
SHS	GPE	0.000	0.000	0.000	0.000	0.110
	MFPE	0.033	0.161	0.344	0.618	1.0
SEPT	GPE	0.000	0.000	0.000	0.000	0.000
	MFPE	0.014	0.014	0.014	0.014	0.014
HNR 15 дБ						
RAPT	GPE	0.000	0.000	0.000	8.12	12.75
	MFPE	0.162	0.271	0.859	2.425	4.814
YIN	GPE	0.000	0.000	0.000	0.000	4.82
	MFPE	0.147	0.238	0.615	1.513	3.101
SWIPE'	GPE	0.000	0.000	0.000	0.000	0.000
	MFPE	0.098	0.201	0.358	0.559	0.977
SHS	GPE	0.000	0.000	0.000	0.057	0.152
	MFPE	0.139	0.226	0.402	0.716	1.53
SEPT	GPE	0.000	0.000	0.000	0.000	0.000
	MFPE	0.019	0.019	0.019	0.019	0.019
HNR 5дБ						
RAPT	GPE	0.000	0.000	0.000	11.31	19.01
	MFPE	0.283	0.482	1.341	3.78	7.514
YIN	GPE	0.000	0.000	0.000	0.000	4.11
	MFPE	0.245	0.349	0.913	2.513	4.101
SWIPE'	GPE	0.000	0.000	0.000	0.000	0.000
	MFPE	0.154	0.28	0.498	0.932	1.89
SHS	GPE	0.000	0.000	0.002	0.103	0.389
	MFPE	0.227	0.32	0.577	1.659	2.734
SEPT	GPE	0.000	0.000	0.000	0.000	0.000
	MFPE	0.020	0.020	0.020	0.020	0.020

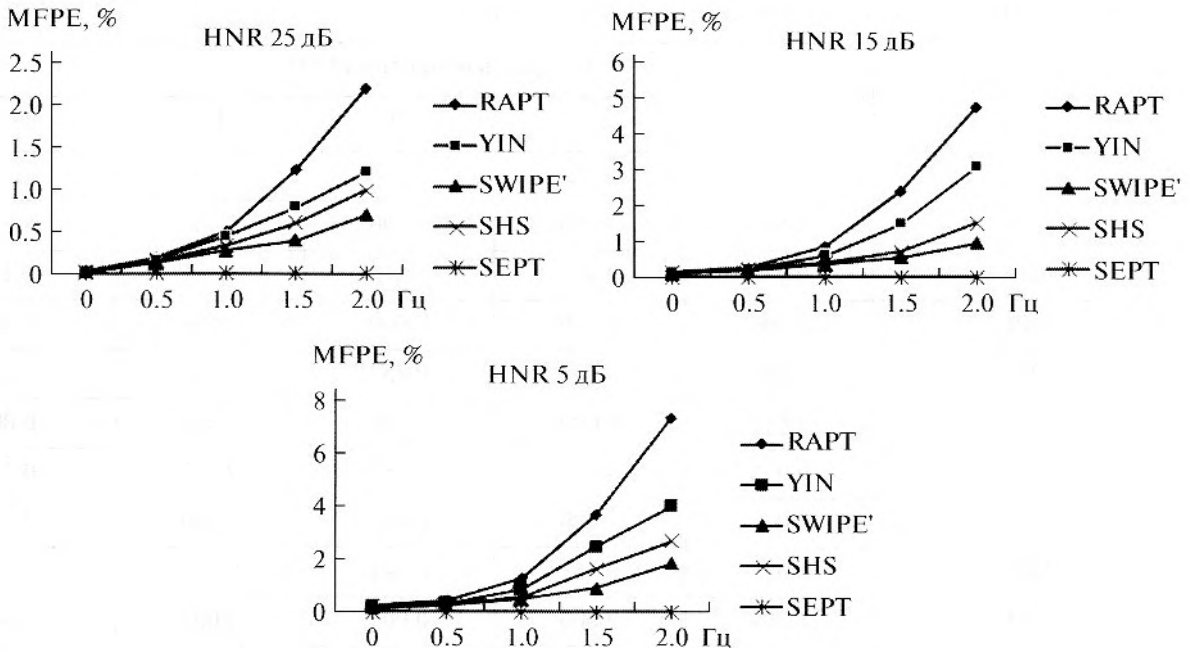

Рис. 7. Изменение точности оценки основного тона с увеличением частотных модуляций.

стотах); $\sqrt{D}$ – среднее квадратичное отклонение σ. Иными словами, нормальная частота основного тона диктора для нижней и верхней границы не превышает 2σ от средней величины. Для конкретного случая (табл. 1) диапазон нижней и верхней границы ЧОТ составляет:

[187.35, 217.858] Гц для "Septv1" и [188.246, 217.714] Гц для "Praat".

Для оценки достоверности использовался набор искусственных синтетических сигналов с заранее известной мгновенной частотой основного тона. Скорость изменения частоты основного тона сигналов изменялась от 0 до 2 Гц/мс. Значения тона находились в пределах от 100 до 350 Гц. Частота дискретизации сигналов – 44.1 кГц. К чистому тональному сигналу добавлялся белый шум различной интенсивности. Интенсивность шума определялась соотношением гармоника/шум (HNR) [5]. Сравнивались алгоритмы SHS, RAPT, YIN, SWIPE' и предлагаемый метод сингулярной оценки ЧОТ (Singular Estimation Pitch Tracking – SEPT) (табл. 2).

Эффективность работы алгоритмов определялась оценками:

1) Процентом грубых ошибок (gross pitch error – GPE)

$$\mathrm{GPE}(\%) = \frac{N_{\mathrm{GPE}}}{N_{\mathrm{V}}} 100,$$

где: N_{GPE} – число фреймов с отклонением полученной оценки более чем на ±20% от настоящего значения основного тона; N_{V} – общее число вокализированных фреймов.

2) Средним процентом мелких ошибок для вокализированных фреймов без грубых ошибок

$$\mathrm{MFPE}(\%) = \frac{1}{N_{\mathrm{FPE}}} \sum_{n=1}^{N_{\mathrm{FPE}}} \frac{\left| F0^{\mathrm{true}}(n) - F0^{\mathrm{est}}(n) \right|}{F0^{\mathrm{true}}(n)} \times 100,$$

где N_{FPE} – число вокализированных фреймов без грубых ошибок; $F0^{\mathrm{true}}(n)$ – истинные значения основного тона; $F0^{\mathrm{est}}(n)$ – оценочные значения основного тона.

Результаты тестирования методов с использованием синтетических сигналов приведены в табл. 2. По полученным результатам можно сделать вывод, что у алгоритмов SHS [1], RAPT [2], YIN [3] и SWIPE' [4] при достаточно небольших изменениях в тональности увеличивается процент грубых и мелких ошибок оценки ЧОТ (табл. 2, рис. 7). Такое ограничение обусловлено используемой периодической моделью речевого сигнала, лежащей в их основе. Указанная модель не допускает изменение периода и амплитуды основного тона на протяжении окна анализа. При этих же условиях сингулярное оценивание ЧОТ оказывает наибольшую робастность к частотным модуляциям, что показывает постоянство оценки MFPE (табл. 2, рис. 7). Таким образом, сингулярный измеритель ЧОТ речевого сигнала учитывает влияние непериодичности, которая имеет место в естественном речевом сигнале. Выполнена сингулярная оценка ЧОТ при непериодичности речевого сигнала (табл. 3). Образцы речи были взяты из базы данных PTDB-TUG [24]. База данных содержит 2342 предложения из корпуса TIMIT, надиктованных дикторами – десятью мужчинами

Таблица 3. Оценка ЧОТ с использованием речевых сигналов

	Мужчины		Женщины		Среднее	
	GPE	MFPE	GPE	MFPE	GPE	MFPE
RAPT	3.701	1.829	5.496	1.237	4.598	1.533
YIN	2.233	1.506	4.276	1.126	3.254	1.316
SWIPE'	0.843	1.305	3.832	0.97	2.337	1.137
SHS	3.214	1.437	4.045	1.09	3.629	1.263
SEPT	0.592	1.201	3.126	0.701	1.859	0.951

и десятью женщинами. База данных включает контрольные сигналы, полученные при помощи ларингографа, и оценочные значения ЧОТ. Анализируемые значения не могут рассматриваться как мгновенные, поэтому нельзя сравнить алгоритмы так же достоверно, как в случае с синтетическими сигналами. Результаты эксперимента с естественной речью показывают, что способ сингулярного оценивания ЧОТ может быть применен для обработки естественных речевых сигналов (табл. 3, рис. 8). Например, у SEPT среднее количество допущенных ошибок GPE и MFPE на 20 и 16% меньше, чем у SWIPE' (табл. 3, рис. 8). Следовательно, предложенный метод оценивания ЧОТ воспроизводит меньшее количество ошибок по сравнению с известными аналогами и способен конкурировать с ними.

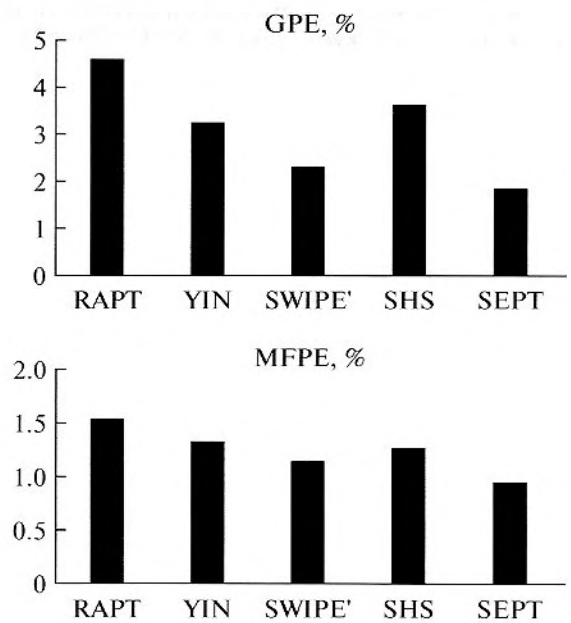

Рис. 8. Средний процент грубых (GPE) и мелких (MFPE) ошибок, воспроизводимых соответствующим алгоритмом оценивания ЧОТ.

ЗАКЛЮЧЕНИЕ

В статье представлена сингулярная модель вокализированного сегмента речевого сигнала, в которой рассмотрена прямая и обратная задачи. Проведено эвристическое построение модели процесса для сингулярного оценивания ЧОТ речи. Представлена численная реализация модели. Рассмотрена оценка эффективности модели. Результаты эксперимента с использованием синтетических сигналов показали, что новая методика измерения ЧОТ решает проблему оценки частотных модуляций основного тона. Результаты эксперимента с естественной речью показывают близкие результаты к другим методам оценивания ЧОТ, что демонстрирует применимость предлагаемого метода оценки к приложениям обработки речевых сигналов. Можно надеяться, что класс анализаторов, основанных на принципе сингулярного спектрального анализа звуковых сигналов, найдет свое широкое применение в акустике, например, для решения задач разделения биоакустических сигналов различной природы [25] или для анализа слабо выраженных сигналов [26] и т.д.

СПИСОК ЛИТЕРАТУРЫ

1. *Hermes D.J.* Measurement of pitch by subharmonic summation // J. Acoust. Soc. Am. 1988. № 83. P. 257–264.
2. *Talkin D.A.* Robust algorithm for tracking (RAPT) // Entropic Research Laboratory Suite 202, 600 Pennsylvania Ave. S.E., U.S.A., Washington D.C. 20003, 1995. P. 495–518.
3. *Cheveigne A., Kawahara H.* YIN, a fundamental frequency estimator for speech and music // J. Acoust. Soc. Am. 2002. V. 111. № 4. P. 1917–1930.
4. *Camacho A., Harris J.G.* A sawtooth waveform inspired pitch estimator for speech and music // J. Acoust. Soc. Am. 2008. V. 123. № 4. P. 1638–1652.
5. *Azarov E., Vashkevich M., Petrovsky A.* Instantaneous pitch estimation based on RAPT framework / Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, Bucharest, 2012. P. 2787–2791.

6. *Бондаренко В.П., Конев А.А., Мещеряков Р.В.* Сегментация и параметрическое описание речевого сигнала // Изв. вузов. Приборостроение. 2007. Т. 50. № 10. С. 3–7.
7. *Golyandina N., Zhigljavsky A.* Singular spectrum analysis for time series. Springer Briefs in Statistics, Springer Berlin Heidelberg, 2013. P. 120.
8. *Данилов Д.Л., Жиглявский А.А.* Главные компоненты временных рядов: метод "Гусеница". Под ред. Данилова Д.Л., Жиглявского А.А. СПб: Пресском, 1997. 308 с.
9. *Gene H.G., Charles F.L.* Matrix computations. 3th edition. U.S.A. Baltimore, Maryland: The Johns Hopkins University Press, 1996. 694 p.
10. *Tony F.C.* An improved algorithm for computing the singular value decomposition // ACM Transaction on Mathematical Software. 1982. V. 8. № 1. P. 72–83.
11. *M. Panju.* Iterative methods for computing eigenvalues and eigenvectors // The Waterloo Mathematics Review. 2011. V. 1. P. 9–18.
12. *Brillinger D.R.* Time series: data analysis and theory. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001. 540 p.
13. *Bagshaw P.C.* Automatic prosodic analysis for computer aided pronunciation teaching. PhD Thesis, Univ. of Edinburgh, Edinburgh, 1994.
14. *Rabiner L.R., Cheng M.J., Rosenberg A.E.* A comparative study of several pitch detection algorithms // IEEE Trans. Acoust. Speech. 1976. № 24. P. 399–423.
15. *Леонов А.С., Сорокин В.Н.*, О точности определения параметров голосового источника // Акуст. журн. 2014. № 60. С. 656–662.
16. *Сорокин В.Н.* Фундаментальные исследования речи и прикладные задачи речевых технологий // Речевые технологии. 2008. № 1. С. 18–49.
17. *Kruskala W.H., Wallisa W.A.* Use of ranks in one-criterion variance analysis // J. American Statistical Association. 1952. V. 47. № 260. P. 583–621.
18. *PRAAT* a computer program: analyze, synthesize, and manipulate speech, and create high-quality pictures for articles and thesis. Phonetic Sciences, Univ. of Amsterdam, 2014. [Электронный ресурс]. Режим доступа URL: http://www.fon.hum.uva.nl/praat (дата обращения 8.12.2014).
19. *Boersma P.* Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound // Proceedings of the Institute of Phonetic Sciences, University of Amsterdam, 1993. V. 17. P. 97–110.
20. *Chaddock R.E* Principles and Methods of Statistics. Houghton Mifflin Comp., 1925. 471 p.
21. *Henriquez P., Alonso J.B., Ferrer M.A., Travieso C.M., Godino-Llorente J.I., Diaz-de-Maria F.* Characterization of healthy and pathological voice through measures based on nonlinear dynamics // IEEE Transaction on audio, speech, and language processing. 2009. V. 17. № 6. P. 1186–1195.
22. *Мещеряков Р.В., Бондаренко В.П.* Диалог как основа построения речевых систем // Кибернетика и системный анализ. 2008. № 2. С. 30–41.
23. *Сорокин В.Н., Макаров И.С.* Определение пола диктора по голосу // Акуст. журн. 2008. Т. 54. № 4. С. 659–668.
24. *Pirker G., Wohlmayr M., Petrik S., Pernkopf F.* A Pitch tracking corpus with evaluation on multipitch tracking scenario // Proceedings of INTERSPEECH. 2011. P. 1509–1512.
25. *Рудницкий А.Г.* Использование метода нелокального усреднения для разделения звуков сердца и звуков дыхания // Акуст. журн. 2014. № 60. С. 688–695.
26. *Малышкин Г.С., Тимофеев В.Н., Туркалова О.И.* Разрешение и оценка параметров слабых сигналов при наличии мешающих источников в зоне Френеля с помощью современных адаптивных алгоритмов // Акуст. журн. 2013. № 59. С. 520–529.