ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 534.784: 621.391

ОБНАРУЖЕНИЕ ПЕРВИЧНЫХ ПРИЗНАКОВ РЕЧЕВОГО СИГНАЛА

канд. техн. наук, доц. И.Б. БУРАЧЁНОК,
д-р техн. наук, проф. В.К. ЖЕЛЕЗНЯК
(Полоцкий государственный университет)

Приведены результаты исследования обнаружения первичных признаков речевого сигнала различными методами на базе гласных фонем мужских голосов средней продолжительностью около 0,25 с в условиях зашумления. Разброс вычисленных значений частоты основного тона рассмотренными методами составил ± 1,37%. Самую низкую относительную погрешность оценки – 0,38%, имеет метод на основе вейвлет-преобразования с использованием в качестве материнского комплексного вейвлета Морле. Относительная погрешность оценки основного тона автокорреляционным методом составила 0,43%. Данный метод предлагается использовать для дальнейших исследований так как он имеет менее сложный алгоритм реализации и не требует больших вычислительных затрат.

Ключевые слова: *речевой сигнал, форманты, основной тон, первичные признаки речевого сигнала, тонкая структура речевого сигнала, маскирующие шумы.*

Введение. Обнаружение слабых сигналов в шумах – это задача, которая имеет отношение ко многим системам связи, в том числе и к системам защиты речевой информации от утечки по техническим каналам. Выделение первичных признаков речевого сигнала, таких как период (или частота) основного тона (ОТ) F_0, Гц, и формант $F_1,\ldots,F_n$, Гц, является необходимым критерием для определения наличия речи в шумах [1].

Из анализа известных методов оценки частоты ОТ (амплитудные методы, основанные на амплитудной селекции; спектральные методы, основанные на частотной селекции; многоканальный пиковый [2]; метод выделения ОТ на основе параллельной обработки [3]; методы на основе кепстрального анализа [3]; автокорреляционные методы [4]; методы на основе линейного предсказания [4]; методы на основе вейвлет-преобразования [5]; метод на основе полигармонической математической модели [6] и др.) следует, что существующие методы определения частоты ОТ исследованы в нормальных условиях и мало исследованы в условиях воздействующих шумов [7]. Поэтому возникла необходимость в проведении дополнительных исследований методов оценки ОТ и первичных признаков элементов речевого сигнала в шумах.

Целью работы является определение метода, повышающего точность оценки информационных признаков речевого сигнала в условиях шумов.

Решаемые задачи:

1. Определение факта присутствия речевого сигнала в анализируемой смеси сигнала и маскирующего его шума выделением частоты основного тона.
2. Определение формант для получения информации об основных характеристиках речевого сигнала и дополнительной информации, характеризующей индивидуальные признаки говорящего.

Спектр речевого сигнала и фонемы русской речи. Энергетический спектр речевого сигнала, представляющий собой область частот, в которой сосредоточена его основная энергия, определяют выражением

$$\beta = 10\log_{10}\left(\frac{G^2(f)}{G_0^2}\right)\Delta f \ ,$$

где $G^2(f)$ – спектральная плотность среднего квадрата звукового давления;

G_0 – порог слышимости, или минимальное звуковое давление на частотах 600…800 Гц, которое начинает ощущаться человеком с нормальным слухом;

$\Delta f = 1$ Гц [9].

В таблице 1 представлены характеристики полос равной разборчивости (с одинаковой вероятностью появления формант в каждой из них) при разбиении спектра речевого сигнала. Выделены критические частотные полосы разборчивости: третья как самая узкая и двадцатая как самая широкая.

Таблица 1. – Разбиение спектра речевого сигнала на двадцать полос равной разборчивости

N_k	1	2	**3**	4	5	6	7	8	9	10
Границы $(f_1 - f_2)$, Гц	100-420	420-570	**570-710**	710-865	865-1030	1030-1220	1220-1410	1410-1600	1600-1780	1780-1960
Центральная частота f_0, Гц	260	495	**640**	788	946	1125	1315	1505	1690	1870
Ширина $2\Delta f$, Гц	320	150	**140**	155	165	190	190	190	180	180

Окончание таблицы 1

N_k	11	12	13	14	15	16	17	18	19	**20**
Границы $(f_1 - f_2)$, Гц	1960-2140	2140-2320	2320-2550	2550-2900	2900-3300	3300-3660	3660-4050	4050-5010	5010-7250	**7250-10000**
Центральная частота f_0, Гц	2050	2230	2435	2725	3100	3480	3855	4530	6130	**8625**
Ширина $2\Delta f$, Гц	180	180	230	350	400	360	390	960	2240	**2750**

Несмотря на то, что речь представляет собой широкополосный процесс, частотный спектр которого сосредоточен в диапазоне от 50…100 Гц до 8000…10000 Гц [8], частоты диапазона 300–3500 Гц приняты МСЭ-Т (Международный союз электросвязи – стандарты, определяющие порядок функционирования и взаимодействия сетей электросвязи) в качестве границ эффективного спектра речи, так как качество речи при ограничении спектра в указанном диапазоне получается вполне удовлетворительным. Согласно исследованиям, представленным в [8], в полосе частот 300–3400 Гц сохраняется удовлетворительная натуральность звучания и слоговая разборчивость составляет около 90%, разборчивость фраз – более 99%.

Звуки русской речи разделяют на гласные и согласные. В русском языке насчитывается 41 звук. Принято выделять 6 гласных (а, о, у, э, и, ы). Гласные звуки составляют примерно 43,5%, а согласные – 56,5% от общего числа звуков, при этом невокализованные звуки составляют 32%. Все гласные звуки являются вокализованными [2; 9; 10].

Гласные звуки речи имеют в среднем длительность около 0,15 с, согласные – около 0,08 с. Для различных людей диапазон звукового давления речевых сигналов достигает 60 дБ. Уровень речи отдельного человека может изменяться в пределах 20–30 дБ. Нормальная речь соответствует акустической мощности 10 мкВт, а шепот – 10^{-3} мкВт [10]. Гласные звуки в среднем имеют мощность 650–700 мкВт, самые слабые согласные – 0,65…0,7 мкВт [10]. Поэтому для обнаружения признаков речевого сигнала традиционно используют алгоритмы, основанные, как правило, на определении характеристик гласных звуков.

В таблице 2 представлены статистические данные вероятности появления звуков в русской речи.

Таблица 2. – Вероятность появления звука в русской речи

Звук русской речи	Вероятность появления в речи	Звук русской речи	Вероятность появления в речи	Звук русской речи	Вероятность появления в речи
О	0,078	Н	0,0476	Г	0,0168
А	0,0713	Р	0,0465	Б	0,0120
Е	0,0626	В	0,0418	Ь	0,0119
И	0,0460	С	0,0398	Й	0,0111
У	0,0226	Т	0,0387	З	0,0104
Я	0,0152	Л	0,0361	Ж	0,0061
Ы	0,0091	К	0,0349	Ц	0,0051
		М	0,0240	Х	0,0049
Пробел/(пауза)	0,17	Д	0,0238	Ф	0,0035
		П	0,0210	Щ	0,0012

Исходя из данных, представленных в таблице 2, в русской речи наиболее часто встречаются гласные звуки «о» и «а», поэтому основные исследования проводились на базе этих гласных фонем мужских и женских голосов средней продолжительностью около 0,25 с (дополнительно рассматривались и некоторые согласные фонемы – «г» и «х» длительностью около 0,10 с) в условиях зашумления при различных отношениях сигнал/шум.

Особенности обнаружения информационных признаков речевого сигнала. Звуки речи являются сложными звуками, так как в процессе речеобразования возникают резонансные явления, собственные частоты которых изменяются в зависимости от произнесенного звука. Дискретные спектры, создаваемые голосовым источником, в первом приближении являются чисто гармоническими. Сформированные звонкие звуки (гласные и согласные) представляют собой *квазипериодическую последовательность треугольной формы*. Импульсы, создаваемые голосовыми связками, кроме основной составляющей содержат большое число гармоник (более 40) [2], причем их амплитуды убывают с увеличением частоты со скоростью приблизительно 12 дБ/октава [8]. Уровень интенсивности гармоник плавно уменьшается с увеличением частоты, т.е. с увеличением их порядкового номера.

По спектральному составу звуки речи различаются числом формант и их расположением в частотном спектре [9]. Отдельным звукам речи может соответствовать до 6 формант, из которых только одна (две) являются определяющими – основными. Под формантой понимают упорядоченную пару значения локального максимума амплитуды и частоты, на котором он достигается [9]. Так как из гласных звуков наибольшей акустической мощностью обладает гласный звук «а», то дальнейшие исследования проводились с использованием данного звука.

На рисунке 1 построена спектрограмма ударного изолированного звука «а», произнесенного мужским голосом. Самые темные области соответствуют локальным максимумам амплитуды и частоты – основным формантам. Вычисленные параметры сигнала: частота основного тона $F_0 = 119$ Гц (мужской голос от 80 до 210 Гц, женский от 150 до 320 Гц); частота первой форманты $F_1 = 601$ Гц; частота второй форманты $F_2 = 1198$ Гц.

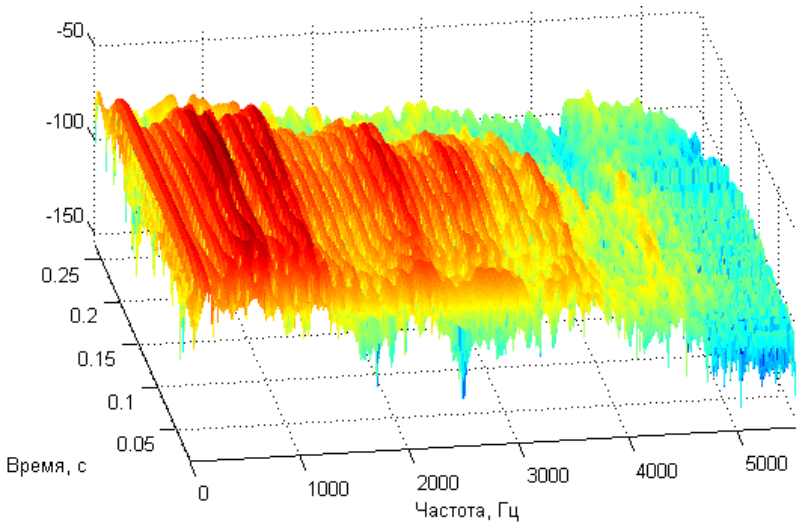


Рисунок 1. – Спектрограмма звука «а»

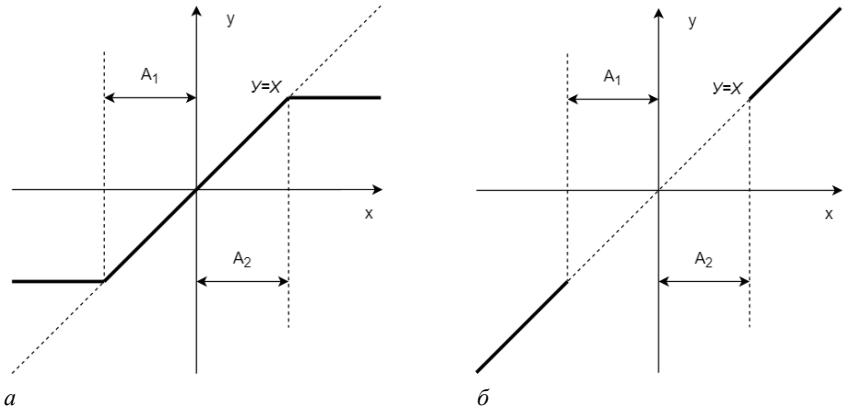
Для узнавания звука важны первые два частотных усиления (или первые две форманты F_1 и F_2). Формант больше, чем две, но именно форманты F_1 и F_2 отвечают за лингвистическую суть любого звука [9]. Если исключить из передачи любую из основных формант, то передаваемый звук исказится.

Спектр звонких звуков в основном расположен в нижней полосе частот (300–3500 Гц) речевого сигнала и сгруппирован вокруг формант, причем скорость изменений формант такова, что спектр остается практически постоянным на промежутках менее 16 мс [10]. Согласно исследованиям основной формантный состав гласных звуков русской речи сосредоточен на следующих частотах: звук «у» – 200…600 Гц, звук «о» – 400…800 Гц, звук «а» – 1000…1400 Гц, звук «и» – 2800…4200 Гц, звук «ы» – 200…600 Гц и 1500…2300 Гц, звук «э» – 600…1000 Гц и 1600…2500 Гц.

Следует отметить и тот факт, что частота импульсов ОТ человеческой речи лежит в пределах от 50–80 Гц (очень низкий голос – бас) до 200–250 Гц (женские и детские голоса) [10]. При произнесении конкретным человеком она непрерывно изменяется (обычно немногим более октавы) в соответствии с ударением, акцентированием звуков и слов, а также их эмоциональной окраской (восклицание, удивление и т.п.). Изменение частоты ОТ называется интонацией. У каждого человека своя интонация, что позволяет идентифицировать говорящего.

В качестве воздействующих факторов применялись белый гауссов шум с распределением по амплитуде и шум ХИП (хаотическая импульсная последовательность) [7]. Также частота ОТ и разборчивость

фонем оценивались при различных уровнях амплитудного ограничения (порогах отсечек) сигнала: клиппировании и центральном ограничении. Клиппирование – ограничение сигнала по амплитуде, т.е. представление речевого сигнала в виде последовательности биполярных прямоугольных импульсов (рисунок 2, *а*). Центральное ограничение речи – искажение переходов через ноль речевого сигнала (рисунок 2, *б*).



а *б*

***а* – ограничение по амплитуде; *б* – центральное ограничение**

Рисунок 2. – Типовые нелинейные искажения воспризводимой речи

Применялось и частотное ограничение спектра до 3500 Гц (целесообразность выбора описана ранее), так как в этом диапазоне расположено подавляющее большинство основных формант звуков речи. Ограничение спектра речевого сигнала позволило получить достаточное качество речи и значительно снизить уровень шума в полосе речевого сигнала. Следует заметить, что качество речи по каналу связи характеризуется рядом факторов: уровнем громкости, не требующим напряжения слуха и голоса; естественным звучанием голоса и низким уровнем помех; достаточной разборчивостью. Безусловно, эти факторы имеют субъективный характер и лишь определяют основные требования к физическим характеристикам речевого сигнала для обнаружения его на фоне шумов.

Исследование первичных признаков речевого сигнала различными методами

Информация об ОТ речевого сигнала очень важна для анализа и синтеза речи, поэтому в первую очередь осуществим анализ существующих методов определения ОТ в условиях шумов.

Метод амплитудной селекции. Как правило, известные методы селекции сводятся к выделению сигналов из шумов путем использования возможных отличий их параметров, таких как несущая частота, ширина спектра, амплитуда, фаза, поляризация и др. Различают частотную, временную, амплитудную, фазовую, поляризационную и пространственную селекции, а также их комбинации.

В результате проведенных исследований методом, основанноым на амплитудной селекции с использованием расстановки меток в точках максимальных значений квазипериодических участков сигнала, можно сделать вывод, что несмотря на то, что данный метод прост в реализации и не требует больших вычислительных ресурсов, результаты его оценки имеют весьма низкую точность и устойчивость даже при небольших уровнях шума. При амплитудном методе в условиях шумов высока вероятность пропуска максимума и неверного определения частоты ОТ.

Спектральный метод. Наиболее известные методы анализа звуков речи основываются на спектральной модели стационарного сигнала

$$S(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} s(t)e^{-i\omega t}dt \ ,$$

где $s(t)$ – массив временных значений речевого сигнала;

$\omega = 2\pi / T_c = 2\pi f_0$ – круговая частота;

t – время.

Как известно, в большинстве языков основная речевая информация (смысловое содержание) передается посредством согласных звуков, а основным недостатком спектрального метода является именно отсутствие характеристик для шумовых составляющих в произносимых согласных звуках.

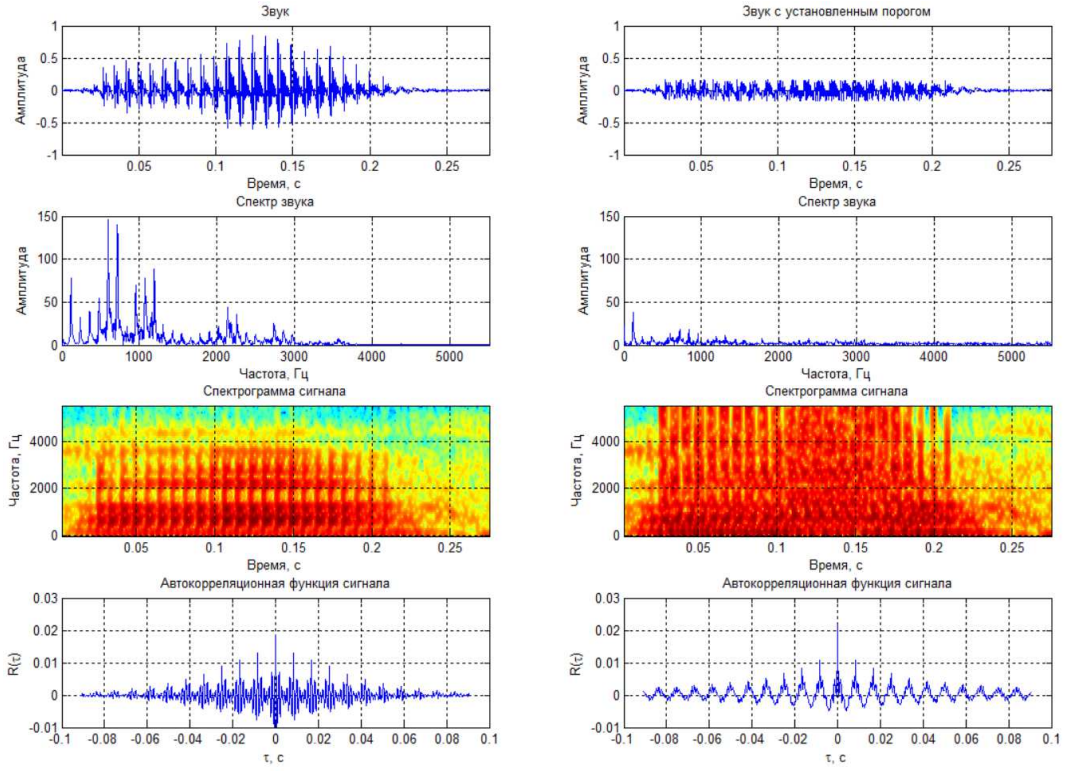
Исследования спектрального метода, основанного на использовании расстановок меток в точках максимальных значений квазипериодических участков частоты, показали, что данный метод имеет самую низкую точность оценки. Поиск максимума необходимо производить в интервале 50–350 Гц, однако в указанной полосе частот часто находится и вторая гармоника, иногда даже с большей энергией, и в этом случае она может быть ошибочно принята за частоту ОТ.

Наилучшее представление о частотно-временной структуре речевого сигнала дает его двумерное или трехмерное изображение (см. рисунок 1). Способы визуализации могут быть различными. Двумерное представление (сонограмма) в осях время–частота позволяет получить картину видимой речи. При этом в процессе произнесения фонемы оттенками цвета изображается перераспределение энергии речевого сигнала между различными частотными полосами. Трехмерное изображение спектра на осях время–частота–амплитуда позволяет осуществить его вращение, что значительно повышает информативность цифрового спектрального анализа.

Автокорреляционный метод. Для анализа речевого сигнала автокорреляционным методом использовалась кратковременная автокорреляционная функция (АКФ):

$$A_s(\tau) = \int_{-\infty}^{\infty} s(t)\, s(t-\tau)\, dt = \int_{-\infty}^{\infty} s(t-\tau)\, s(t)\, dt = A_s(-\tau) \ .$$

АКФ представляет собой функцию скалярного произведения сигнала и его копии, сдвинутой на интервал τ, при $-\infty < \tau < \infty$, однако с учетом четности вычисление АКФ чаще производится только для положительных значений τ. Это обычно используют на практике, когда сигналы задаются на интервале положительных значений аргументов от 0 до T, что дает возможность продления интервала нулевыми значениями, если это необходимо для математических операций. В этих границах вычислений более удобным является сдвиг копии сигнала влево по оси аргументов. Для вокализованных фонем АКФ имеет четкий максимум в районе задержек, равный периоду ОТ. На рисунке 3 показаны возможности оценки основного тона амплитудным, спектральным и автокорреляционным методами, а также совместное использование амплитудного и спектрального методов при построении сонограммы.



а *б*

***а* – исходные характеристики сигнала;**
***б* – характеристики сигнала, искаженного установлением порога отсечки**

Рисунок 3. – Пример искаженного сигнала с установлением порога отсечки

Для повышения точности оценки значений частоты основного тона гласных звуков определим экстремумы АКФ-сигнала, используя математический метод. Найдем производную данной функции с целью определения положения точек, в которых данная производная равна нулю (рисунок 4, *а*):

$$\frac{\partial R_v(\tau,0)}{\partial \tau} = \int_{-\infty}^{\infty} s_1(t) \frac{\partial}{\partial \tau}\left[s_1^*(t-\tau)\right] dt \ .$$

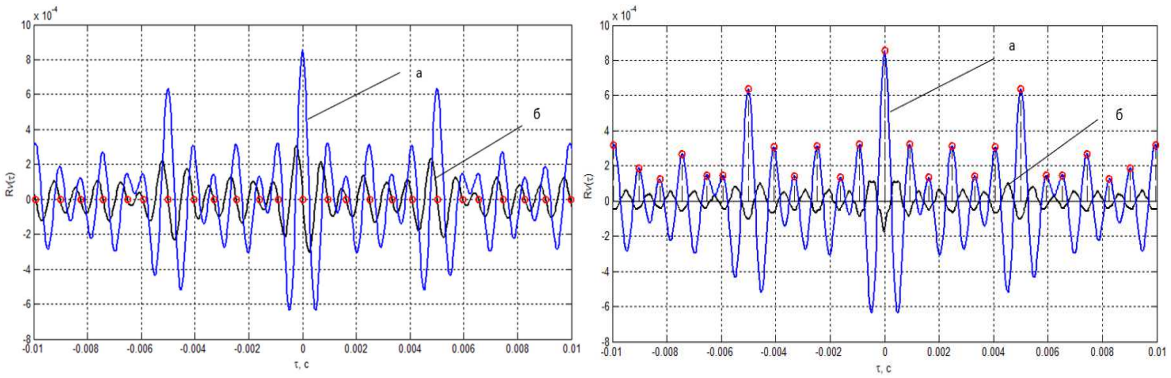
На рисунке 4, *а* показаны графики функций $R_v(\tau,0)$ и $\frac{\partial R_v(\tau,0)}{\partial\tau}$, показывающие, что нули производной АКФ соответствуют ее экстремумам. В точке максимального совпадения производная от АКФ имеет *S*-образную форму, пересекающую нулевой уровень.

Далее определяем, какие из найденных точек являются точками минимума, а какие – максимума. Тип экстремума (максимум или минимум) определяется знаком второй производной в этой точке. Для этого находим значения второй производной согласно выражению

$$\frac{\partial^2 R_v(\tau,0)}{\partial\tau^2} = \int_{-\infty}^{\infty} s_1(t)\frac{\partial^2}{\partial\tau^2}\left[s_1^*(t-\tau)\right]dt$$

и определяем ее знак.

На рисунке 4, *б* показаны графики функций $R_v(\tau,0)$ и $\frac{\partial^2 R_v(\tau,0)}{\partial\tau^2}$. Расстояние между двумя максимумами есть время основного тона, обратная величина которого равна частоте основного тона.



а *б*

***а* – совместное отображение АКФ и ее первой производной речевого сигнала:**
а – АКФ, б – первая производная от АКФ;
***б* – совместное отображение АКФ и ее второй производной речевого сигнала:**
а –АКФ, б – вторая производная от АКФ

Рисунок 4. – Демонстрация процесса повышения точности оценки основного тона автокорреляционным методом

При различных уровнях амплитудного ограничения сигнала речи (порогах отсечек) – клиппировании – согласные звуки произносятся с шипением, что снижает их разборчивость, также ухудшается разборчивость и гласных звуков. При порогах отсечки сигнала выше 35–40% данный расчет не представляется возможным. При центральном амплитудном ограничении сигнала речи (нулевой отсечке) происходят самые существенные изменения сигнала: «съедаются» некоторые звуки, голос становится более высоким, появляются хрипы, повышается уровень шумов. Относительная погрешность оценки основного тона автокорреляционным методом составила 0,43%.

Кепстральный метод. Разработанный Р.В. Шафером и Л.Р. Рабинером метод кепстрального анализа основан на вычислении и анализе кепстра – обратного преобразования Фурье логарифма спектра мощности сигнала [3]:

$$C_s(q) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \ln\left|S(\omega)\right|^2 e^{i\omega q}d\omega\,,$$

где $S(\omega)$ – спектр сигнала.

Одной из особенностей данного метода является усиление влияния низкочастотных компонент шума за счет операции логарифмирования спектра. Наличие выброса в кепстре в диапазоне 3–20 мс очень точно указывает на то, что данный сегмент является вокализованным. Наилучший результат этот метод дает при оценке вокализованных звуков [3]. Однако работа ведется в нереальном масштабе времени, а для повышения точности оценки необходимо применять временные окна и операции сглаживания. Следовательно, из-за невысокой стойкости к шумам и вычислительной сложности кепстральный подход не позволяет получить хорошие результаты для решения задач выделения ОТ в шумах.

Метод линейного предсказания. Применение линейного предсказания LPC (Linear Predictive Coding) описано в работах Итакура и Санто, Атан и Ханауэр [4]. Дж.Д. Маркел для автоматического оценивания формантных траекторий с использованием линейного предсказания предложил упрощенную процедуру. Оценку спектра сигнала на выходе линейного тракта с неизвестными параметрами можно представить выражением [3]

$$H(z) = \frac{\sum_{i=0}^{N-1} b_i z^{-1}}{1 + \sum_{k=1}^{M-1} a_k z^{-1}}$$

или соответствующим разностным уравнением

$$y(n) = \sum_{i=0}^{N-1} b_i x(n-i) - \sum_{k=0}^{M-1} a_k y(n-k),$$

где на некоторую линейную модель системы с передаточной функцией $H(z)$ воздействует сигнал возбуждения $x(n)$, а на ее выходе формируется сигнал $y(n)$.

Применение линейного предсказания открывает широкие возможности в области компрессии речевых данных. Однако в наших условиях при частотах ОТ выше 200 Гц данный метод оценки приводит к плохим результатам.

Метод на основе вейвлет-преобразования. Это относительно новый, развивающийся метод оценки частоты ОТ при использовании амплитудной вейвлет-функции

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(t)\, \psi^* \left(\frac{t-b}{a} \right) dt, \qquad a,b \in R,\ a > 0,$$

где a – масштаб;

b – сдвиг;

ψ – базисная функция.

Речевой сигнал (РС) – это нестационарный процесс, в котором информативным является сам факт изменения его частотно-временных характеристик. Для выполнения анализа таких процессов требуются специальные базисные функции, имеющие способность одновременно выявлять в анализируемом сигнале как его частотные, так и временные характеристики. Другими словами, сами функции должны обладать свойствами частотно-временной локализации. Вейвлет-преобразование имеет эти свойства, а также ряд существенных преимуществ при выполнении высокоточного анализа сложных нестационарных сигналов.

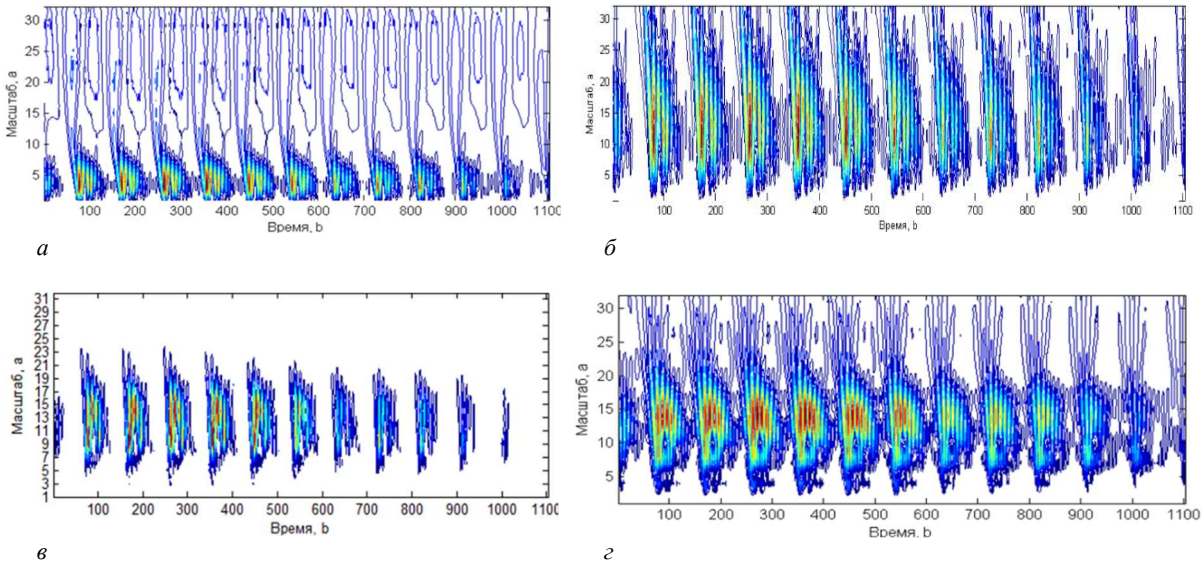
Для исследований был выбран участок сигнала длительностью 0,1 с, где присутствует наибольшая интенсивность звука и сосредоточены основные форманты F_1 и F_2. Исследуемый фрагмент гласного звука «а» представим временным рядом со значениями функции, следующими друг за другом с постоянным интервалом времени Δt: $s_k = s(t_k)$, $t_k = \Delta t k$, $k = 0,1,...N-1$. Интервал временного окна, где процесс можно считать квазистационарным, определим равным $\Delta t \le 20$ мс.

Оценка локального спектра энергии (анализ распределения энергии по частотно-временной шкале) осуществлялась на основании построения скалограммы, которая для дискретного сигнала может быть записана как $S(a_i,b_j) = \left| W_A(a_i,b_j) \right|^2$, в трехмерном пространстве координат (a,b,S) и представления в виде топологической карты при изображении поверхности $S(a,b)$ в координатах (a,b).

Скалограммы заданного участка фонемы «а» (см. рисунок 1), представляющие масштабное распределение энергии сигнала, где по оси времени отложена величина сдвига вейвлет-функции – b, по оси ординат – масштаб a, с использованием различных вейвлетов при одинаковых значениях масштабных коэффициентов представлены на рисунке 5. Рассмотрены вейвлеты: «мексиканская шляпа» (см. рисунок 5, *а*); гауссова типа (см. рисунок 5, *б*); Мейера (см. рисунок 5, *в*) и комплексный вейвлет Морле (см. рисунок 5, *г*). Применение комплексного вейвлета Морле в качестве материнского при одних тех же значениях масштабных коэффициентов a и b обладает большей детализаций и информативностью.

В работе [11] показана возможность получения тонкой структуры речевого сигнала в случае применения комплексного вейвлета Морле. Изображение его вейвлет-скалограммы выявляет наличие разномасштабной периодичности, содержащейся в анализируемых зависимостях, показывая наличие появившихся частотных составляющих, не соответствующих собственным частотам рассматриваемого РС. Комплексный вейвлет Морле также имеет близкое сходство с речевыми фрагментами (подобен импульсным составляющим нестационарных сигналов) и обладает лучшей по сравнению с другими базисами частотной локализацией, он имеет более узкий Фурье-образ и продолжителен во временной области. Присутствие доминирующей частоты позволяет варьировать его избирательностью в частотной области [12]. Из рисунка 5

следует, что гармоники основного тона остаются устойчивыми на протяжении всего временного промежутка, в то время как выявленные гармоники высших порядков со временем затухают.



***а* – вейвлет «мексиканская шляпа»; *б* – вейвлет гауссова типа;**
***в* – вейвлет Мейера; *г* – комплексный вейвлет Морле**

Рисунок 5. – Примеры скалограмм звука «а», полученных с использованием различных материнских вейвлетов при одинаково заданных значениях масштабных коэффициентах *a* и *b*

Можно отсечь влияние контуров, выделив те точки скалограммы, в которых она имеет максимумы (локальные экстремумы) по переменным a и b, построив так называемый скелетон (sceleton) (рисунок 6), выявляющий структуру анализируемого сигнала

$$S_c(a_i,b_j)=\begin{cases} S_{ij}, \text{ если } S_{i-1,j}<S_{i,j}>S_{i+1,j} \\ \text{или } S_{i,j-1}<S_{i,j}>S_{i,j+1} \\ 0, \text{ в остальных случаях.} \end{cases}$$

Увеличим масштаб скелетона исследуемой фонемы «а» (см. рисунок 5, *г*). Его фрагмент представлен на рисунке 6.

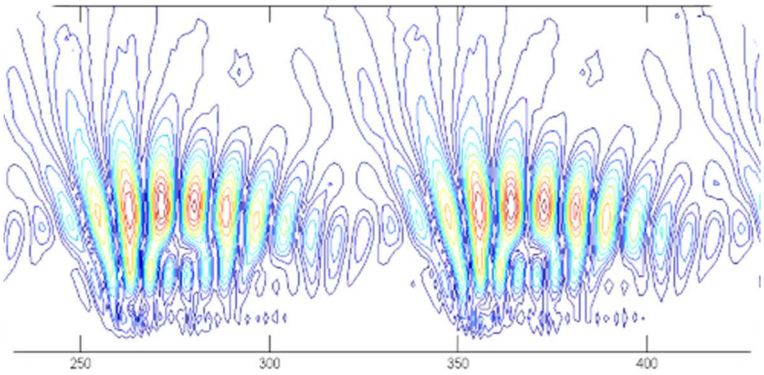
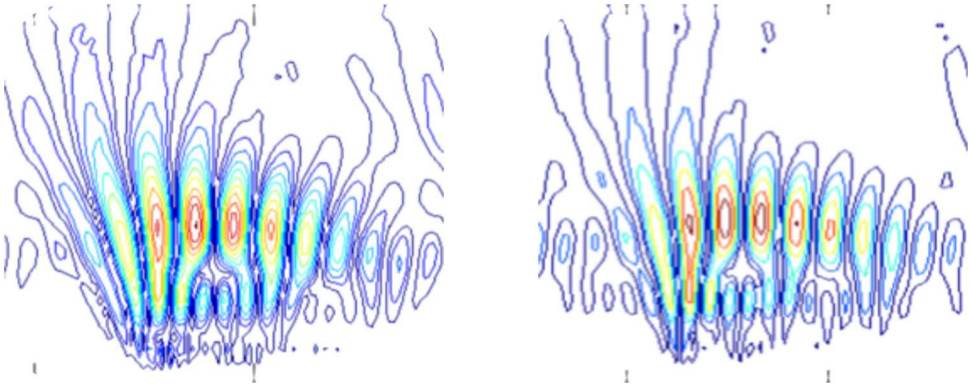


Рисунок 6. – Фрагмент скелетона фонемы «а» (два рядом идущих скейла), полученный с использованием комплексного вейвлета Морле

Построение изолиний (изоуровней) позволяет получать картины линий локальных экстремумов, тем самым, четко визуализировать структуру анализируемого процесса, а также отслеживать изменения интенсивности амплитуд вейвлет-преобразования на разных масштабах и во времени. Цветовая гамма наглядно демонстрирует зависимость интенсивности от изменения вейвлет-коэффициента S. На рисунке 6 темно-красные изолинии соответствуют положительным, а темно-синие – отрицательным значениям $W(a,b)$. Ясно, что значение амплитуды вейвлет-преобразования в точке (a_0,b_0) тем больше (по абсолютной величине),

чем сильнее корреляция между вейвлетом данного масштаба и поведением сигнала в окрестности $t = b_0$. Картина коэффициентов демонстрирует, что процесс составляют компоненты разных масштабов: экстремумы $W(a,b)$ наблюдаются на разных масштабах, интенсивность их меняется и со временем, и с масштабом.

На рисунке 7 показаны фрагменты скелетонов фонемы «а», произнесенной одним человеком в разных условиях. Визуальное сходство представленных фрагментов демонстрирует возможность осуществления идентификации дикторов для фонограмм, записанных в каналах с высоким уровнем помех и искажений, а также фонограмм с произвольной устной речью дикторов, находящихся в различных психофизиологических состояниях.

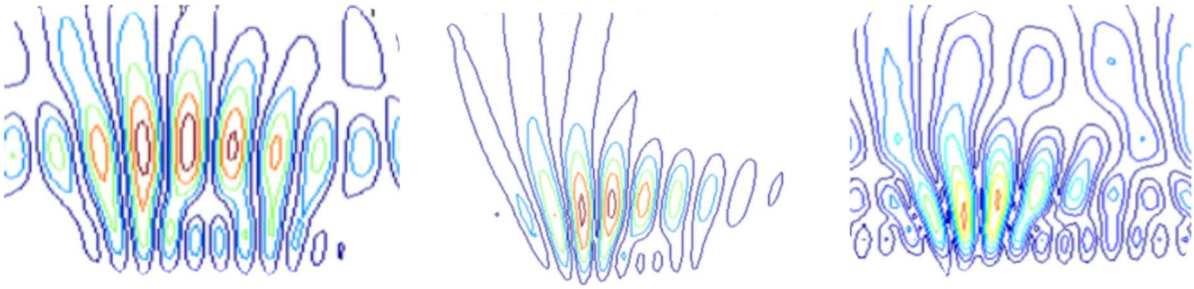


а *б*

***а* – отдельно произнесенный ударный звук; *б* – вырезанный звук из слова «барабан»**

Рисунок 7. – Фрагмент фонемы «а», произнесенной одним человеком в разных условиях

На рисунке 8 приведены фрагменты фонемы «а», произнесенной разными людьми (мужские голоса) с различной частотой основного тона F_0 в одинаковых условиях.



а *б* *в*

***а* – F_0 = 111 Гц; *б* – F_0 = 96 Гц; *в* – F_0 = 152 Гц**

Рисунок 8. – Фрагмент фонемы «а», произнесенной разными людьми (мужские голоса) с различной частотой основного тона F_0 в одинаковых условиях

Представленные фрагменты фонемы «а» (см. рисунок 8) демонстрируют наличие отличительных особенностей тонкой структуры отдельно взятого звука речи.

Огромным преимуществом метода на основе вейвлет-преобразования перед всеми рассмотренными методами является то, что применение данного вида анализа позволяет выделить индивидуальные признаки говорящего и проводить фоноскопическую экспертизу фонем речи на предмет идентификации по голосу, а также выявлять признаки монтажа или иных изменений, вносимых в аудиозапись. Однако несмотря на то, что у данного метода самая низкая относительная погрешность оценки ОТ (0,38%), он имеет сложный алгоритм и требует больших вычислительных затрат.

Дискретно-квантованное представление речевых сигналов. В отличие от непрерывного исходного сигнала его дискретное представление позволяет осуществить передачу на большие расстояния при высокой скорости и качестве передаваемого сигнала. Погрешность восстановления исходного сигнала зависит от вида исходной функции; процесса квантования, связанного с округлением значений непрерывного сигнала; интервала квантования и алгоритма восстановления. Квантование сигналов по амплитуде позволяет эффективно подавлять помехи, если только среднеквадратическое значение помех мало по сравнению с разностью между дискретными уровнями [13].

По результатам исследований установлено, что отношение сигнал/шум принятого сигнала существенно зависит от шага квантования. Разряд квантователя заметно влияет и на качество речи. Например, при использовании 7-разрядного аналого-цифрового преобразователя (АЦП) – квантующего устройства – заметных изменений не происходит, наблюдается лишь появление незначительного шума, однако если мы берем 4-разрядный квантователь, то речь по-прежнему остается разборчивой, но уже сопровождается треском, а при обработке 3-разрядным квантователем разобрать речь практически невозможно.

Следует заметить и то, что чем меньше интервал дискретизации (чем выше частота дискретизации), тем точнее отображается исходная функция (форма восстановленного сигнала приближается к оригиналу), и тем меньше ошибки квантования по времени. Однако не следует забывать, что при этом увеличивается количество обрабатываемой информации, что требует увеличения как объема памяти, так и быстродействия устройства обработки информации. На практике частота дискретизации выбирается исходя из теоремы Котельникова и для речевого сигнала составляет не менее 8 кГц, поэтому в дальнейших исследованиях для передачи сигналов рекомендуется использование 7 бит, или числа уровней квантования $2^7 = 128$ (шаг квантования – 7,8 мВ), частоты дискретизации $F_д = 192$ кГц (период дискретизации – 5,2 мкс) посредством применения универсального АЦП/ЦАП LCard E14-440D (14 бит, 400 кГц, для подключения 16 дифференциальных каналов или 32 с «общей землей») с интерфейсом USB 2.0. Диапазон входных значений находится в пределах от –10 до 10 В. Разрешение двоичного АЦП по напряжению составляет $\left(10-\left(-10\right)\right)/16384 = 20/16384 = 0{,}00122$ В $= 1{,}22$ мВ. Эффективная разрядность – 13,2 бит. Ошибка квантования $1/16384 = 6{,}1\times10-5 = 0{,}0061\%$. Используемый АЦП/ЦАП удобен для создания портативных измерительных систем на базе ноутбука, так как не требует дополнительного источника питания.

Заключение. При разговоре, во время перехода от гласных звуков к согласным и наоборот, частота ОТ меняется в значительных пределах. Среднее квадратическое отклонение частоты ОТ, произнесенного одним человеком, для мужского и женского голосов соответственно равно 17 и 27 Гц. Разброс вычисленных значений частоты ОТ рассмотренными методами составил ± 1,37% [1].

На основании обработки экспериментальных данных наилучшие результаты оценки ОТ при внесении искажений в сигнал различными отсечками – клиппированием и нулевой отсечкой, а также в условиях шумов получены автокорреляционным методом, позволяющим оценить периодичность сигнала в зависимости от его задержки. Данный метод не требует больших вычислительных ресурсов, позволяет наиболее точно определить частоту ОТ (относительная погрешность оценки 0,43%) произнесенных фонем при различных уровнях амплитудного ограничения сигналов и обнаруживает речевой сигнал даже на фоне мощных шумов.

Применение вейвлет-преобразования позволяет проводить оценку скейлинговых параметров сигналов и с высокой точностью определять значение частоты ОТ (относительная погрешность оценки не превышает 0,38%). Выбор комплексного вейвлета Морле в качестве материнского и результаты анализа экспериментальных исследований параметров гласных звуков, полученные с помощью данного вейвлет-преобразования, позволят в дальнейшем в силу масштабирующих свойств вейвлетного базиса, обеспечивающего связь между частотным и временным разрешением, гибко управлять настройкой параметров вейвлета для получения тонкой структуры сигнала конкретного диктора в реальном масштабе времени, а также решать задачи очистки сигнала от шума при цифровой обработке сигналов.

ЛИТЕРАТУРА

1. Бурачёнок, И.Б. Оценка тонкой структуры информационных признаков речевого сигнала / И.Б. Бурачёнок, В.К. Железняк // Современные средства связи : материалы XVIII Междунар. науч.-техн. конф., Минск, 15–16 окт. 2013 г. / Белорус. гос. акад. связи; редкол.: А.О. Зеневич [и др.]. – Минск, 2013. – С. 184–186.
2. Михайлов, В.Г. Измерение параметров речи / В.Г. Михайлов, Л.В. Златоустова ; под ред. М.А. Сапожкова. – М. : Радио и связь, 1987 – 168 с.
3. Rabiner, L.R. Digital processing of speech signals / Lawrence R. Rabiner, Ronald W. Schafer. – New Jersey : Prentice Hall PTR, Englewood Cliffs, 1978. – 512 p.
4. Markel, J.D. Linear Prediction of Speech, Springer-Verlag / J.D. Markel, A.H. Gray, Jr. – New York, 1976.
5. Рассказова, С.И. Метод формантного анализа на основе вейвлет-преобразования в системах распознавания речи / С.И. Рассказова, А.И. Власов // Наукоемкие технологии и интеллектуальные системы: сб. тр. IX Науч.-техн. конф. – М. : МГТУ им. Н.Э. Баумана, 2007. – С. 38–43.
6. Голубинский, А.Н. Расчет частоты основного тона речевого сигнала на основе полигармонической математической модели / А.Н. Голубинский. // Вестн. Воронеж. ин-та МВД России. – 2009. – № 1. – С. 81–90.
7. Бурачёнок, И.Б. Обнаружение измерительных сигналов в маскирующих шумах высокого уровня / И.Б. Бурачёнок, В.К. Железняк, А.Г. Филиппович // Вест. Полоц. гос. ун-та. Сер. С, Фундам. науки. – 2018. – № 4. – С. 2–9.
8. Гитлиц, М.В. Теоретические основы многоканальной связи : учеб. пособие для вузов связи / М.В. Гитлиц, А.Ю. Лев – М. : Радио и связь, 1985. – 248 с.
9. Фланаган, Дж. Анализ, синтез и восприятие речи / Дж. Фланаган ; пер. с англ. под ред. А.А. Пирогова. – М. : Связь, 1968. – 396 с.

10. Стопцов Н.А. Связь под водой : справочное пособие / Н.А. Стопцов, В.И. Бойцов, В.Н. Шелемин. – Л. : Судостроение, 1990. – 248 с.
11. Бурачёнок, И.Б. Анализ вейвлет-преобразованием тонкой структуры гласных звуков речевого сигнала / И.Б. Бурачёнок, В.К. Железняк // Теоретические и прикладные аспекты информационной безопасности : материалы междунар. науч.-практ. конф., Минск, 19 июня 2014 г. / Акад. М-ва внутр. дел Респ. Беларусь ; редкол.: В.Б. Шабанов (отв. ред.) [и др.]. – Минск, 2015. – С. 124–128.
12. Бурачёнок, И.Б. Вейвлет преобразования при оценке защищенности речевого сигнала в условиях наличия шумов [Электронный ресурс] / И.Б. Бурачёнок, В.К. Железняк // Информационно-коммуникационные технологии: достижения, проблемы, инновации (ИКТ-2018) : электрон. сб. ст. I Междунар. науч.-практ. конф., посвященной 50-летию Полоцкого государственного университета, Новополоцк, 14–15 июня 2018 г. / Полоц. гос. ун-т. – Новополоцк, 2018. –383 с. – С. 323–327. – 1 электрон. опт. диск (CD-ROM).
13. Шелухин, О.И. Цифровая обработка и передача речи / Шелухин О.И., Лукьянцев Н.Ф. ; под ред. О.И. Шелухина. – М. : Радио и связь, 2000. – 456 с.
14. Скучик, Е. Основы акустики / Е. Скучик. – М. : Мир, – Т. 1.– 1976. – 520 с.
15. Передача информации в подвижных системах связи / В.Ю. Бабков [и др.]. – СПб. : СПбГУТ, 1999. – 152 с.

Поступила 27.08.2020

DETECTION OF PRIMARY SIGNS OF A SPEECH SIGNAL

I. BURACHONAK, V. ZHELEZNYAK

The results of a study of the detection of primary signs of a speech signal by various methods based on vowel phonemes of male voices with an average duration of about 0.25 s in noisy conditions are presented. The spread of the calculated values of the pitch frequency discussed methods was ± 1,37%. The method based on the wavelet transform with the use of the complex Morlet wavelet as the parent has the lowest relative estimation error – 0.38%. Relative error autocorrelation pitch estimation method was 0.43%. This method is proposed to be used for further research as it has a less complex implementation algorithm and does not require large computational costs.

Keywords: *speech signal, formants, pitch frequency, primary signs of a speech signal, the fine structure of the speech signal, masking noise.*