УДК 004.934:681.518

АКУСТИЧЕСКИЕ ДЕСКРИПТОРЫ ГАРМОНИЧЕСКОЙ СТРУКТУРЫ РЕЧИ ДЛЯ ОЦЕНКИ ЭМОЦИЙ

О. В. Мельник, д.т.н., профессор кафедры ИИБМТ, Рязань, Россия;
orcid.org/0000-0002-3513-2180, e-mail: omela111@yandex.ru
С. И. Бабаев, к.т.н., доцент кафедры ЭВМ, Рязань, Россия;
orcid.org/ 0000-0001-5829-8223, e-mail: babaev.s.i@gmail.com
М. Н. Сараев, аспирант РГРТУ, Рязань, Россия;
orcid.org/0009-0006-5118-3478, e-mail: mixailr@mail.ru

Представлены классические акустические дескрипторы, основанные на гармонической структуре речи, применяемые для автоматической оценки эмоциональных состояний (нейтральное состояние – стресс). ***Цель работы*** *– систематизировать методы анализа гармонической структуры речи, раскрыть их физиологические основания и оценить информативность в отношении эмоциональных изменений. Рассмотрены ключевые методы: анализ отношения гармоники к шуму (Harmonic-to-Noise Ratio, HNR), оценка основной частоты тона (Fundamental Frequency, F_0), параметры нестабильности периода и амплитуды: джиттер (Jitter) и шиммер (Shimmer), спектральный анализ на основе коротковременного преобразования Фурье (Short-Time Fourier Transform, STFT), кепстральный анализ (Cepstral Analysis), формантный анализ (Formant Analysis). Описаны их алгоритмы и чувствительность к эмоциональным изменениям. Особый фокус сделан на физиологически интерпретируемых параметрах (F_0, HNR, Jitter, Shimmer) и лежащих в основе их вычисления фундаментальных методах – спектральном и кепстральном анализе. Отмечены ограничения каждого метода, и даны рекомендации по выбору дескрипторов. Практическая значимость рассмотренных методов заключается в демонстрации их рабочей применимости на иллюстративном материале: в паре записей (нейтральное состояние – стресс) зафиксированы характерные изменения – снижение HNR, рост Jitter и Shimmer, увеличение энергии сигнала ($MFCC_0$), а также возрастание вариабельности формант (F_1 – F_4), что подтверждает чувствительность дескрипторов к эмоциональному напряжению и обосновывает использование комбинированного набора признаков. Статья будет полезна специалистам в области обработки сигналов, психолингвистики и систем распознавания эмоций.*

Ключевые слова: *гармоники речи, анализ отношения гармоника/шум, оценка основной частоты тона, джиттер, шиммер, спектральный анализ, формантный анализ, кепстральный анализ, оценка эмоций, стресс.*

DOI: 10.21667/1995-4565-2026-95-171-185

Введение

Эмоциональный анализ речи представляет собой динамично развивающееся междисциплинарное направление на стыке лингвистики, психологии и компьютерных наук [1]. В клинической практике голосовые маркеры обеспечивают объективную диагностику депрессивных и тревожных расстройств на ранних стадиях [2], что особенно востребовано в телемедицине. В системах человеко-машинного взаимодействия – от виртуальных ассистентов до социальных роботов – распознавание эмоций по голосу значительно повышает качество коммуникации и пользовательский опыт [3, 4].

Современный этап развития характеризуется переходом к методам глубокого обучения, однако интерпретируемость результатов остается критическим требованием для клинических применений [5]. Классические акустические параметры гармонической структуры речи (F_0, HNR, Jitter и Shimmer) обеспечивают физиологически обоснованную интерпретацию, отражая физиологические изменения в голосовом аппарате при

эмоциональном возбуждении [6], что делает их незаменимыми для создания объяснимых систем анализа.

Систематизация и анализ эффективности классических дескрипторов гармонической структуры речи представляют актуальную задачу для формирования научно обоснованного подхода к разработке робастных и интерпретируемых систем распознавания эмоций, что определяет практическую значимость данного исследования для специалистов в области обработки сигналов, психолингвистики и систем распознавания эмоций.

Несмотря на явные преимущества классических акустических дескрипторов гармонической структуры речи в плане интерпретируемости и связи с физиологией голосообразования, их применение для надежной и робастной оценки эмоций, особенно в условиях стресса, сталкивается с рядом серьезных вызовов. Ключевая проблема заключается в противоречии между физиологической обоснованностью этих параметров и сложностью установления однозначных, устойчивых и универсальных закономерностей их изменения под влиянием разнообразных эмоциональных состояний в реальных условиях [1, 7].

Это усугубляется методологической разрозненностью: отсутствием стандартизированных подходов к расчету и интерпретации параметров (например, вариативность алгоритмов оценки Jitter и Shimmer) [8], высокой чувствительностью многих дескрипторов к артефактам записи и обработки (шум, тип микрофона, алгоритмы предобработки) [9], а также существенным межиндивидуальным и межкультурным варьированием голосовых характеристик, затрудняющим создание универсальных моделей [10].

Постановка задачи

Цель работы – систематизировать методы анализа гармонической структуры речи (F_0, HNR, Jitter, Shimmer, Formant Analysis, Cepstral Analysis), раскрыть их физиологические основания и оценить информативность в отношении эмоциональных изменений (по шкале нейтральное состояние – стресс»).

Для достижения поставленной цели решаются *следующие задачи*.

1. Объединить и структурировать сведения о шести ключевых акустических дескрипторах: F_0, HNR, Jitter, Shimmer, Formant Analysis, Cepstral Analysis.

2. Проиллюстрировать связь каждого дескриптора с механизмами голосообразования и физиологическими изменениями в голосовом тракте при эмоциональных возмущениях.

Теоретические исследования

Механизмы голосообразования и физиологии речи

Голосовой тракт представляет собой комплекс анатомических структур, отвечающих за формирование речевых звуков [11]. Он состоит из трех основных сегментов.

1. *Дыхательный аппарат* (источник энергии). Легкие и диафрагма создают субглоттальное давление, обеспечивая контролируемый поток воздуха, необходимый для фонации. Речевое дыхание характеризуется активным и продолжительным выдохом [12].

2. *Гортань* (источник звука). Содержит голосовые складки, колебания которых генерируют исходный звук. Степень и регулярность смыкания складок, их натяжение и масса напрямую влияют на ключевые акустические параметры: частоту основного тона (F_0), отношение гармоники к шуму (HNR), джиттер и шиммер [13].

3. *Резонаторные полости* (фильтр-модулятор). Глотка, ротовая и носовая полости. Их форма, управляемая артикуляторами (язык, губы, нижняя челюсть, мягкое небо), изменяет спектральный состав звука, формируя характерные спектральные максимумы – форманты (F_1, F_2, F_3 и т.д.), определяющие тембр и артикуляцию [1, 7, 13].

Процесс фонации – это механизм создания основного тона голоса, базирующийся на циклическом смыкании и размыкании голосовых складок под действием субглоттального давления:

– *выдох*: поток воздуха из легких встречает сомкнутые складки;

– *давление снизу размыкает складки*, позволяя воздушной струе пройти;

– *эффект Бернулли* (снижение давления в быстром потоке) и эластичность тканей приводят к обратному смыканию складок снизу вверх [12, 13].

Этот цикл повторяется с высокой частотой, порождая периодическую вибрацию. Частота этих колебаний определяет основную частоту тона (F_0), а амплитуда – громкость звука. Форма и объем резонаторных полостей модулируют гармоники исходного сигнала, задавая его тембр за счет формант. Нормальные значения F_0 у человека составляют: для мужчин 85 – 180 Гц, для женщин 165 – 255 Гц, для детей 250 – 400 Гц, достигая у профессиональных певцов 60 – 1000 Гц [11].

Формирование гармонической структуры речи и ее изменение под влиянием эмоционального стресса

Голосовые складки генерируют периодический сигнал, состоящий из основной частоты (F_0) и ее гармоник. Резонаторные полости (глотка, ротовая и носовая полости) действуют как фильтры, формируя спектральную огибающую и форманты (F_1, F_2, F_3) – локальные пики энергии в спектре [7]. Озвученные (voiced) сегменты речи возникают при вибрации голосовых складок, они несут основную эмоциональную информацию благодаря четкой гармонической структуре [14]. Неозвученные (unvoiced) сегменты, напротив, формируются турбулентным потоком воздуха без вибрации голосовых складок и характеризуются апериодическим шумом [8]. Классификация речевых звуков по типу фонации приведена в таблице 1.

Таблица 1 – Классификация речевых звуков по типу фонации
Table 1 – Classification of speech sounds by type of phonation

Тип звука	Физиологический механизм	Акустические свойства	Примеры
Озвученные (voiced)	Вибрация голосовых складок	Периодическийсигнал, F_0 и гармоники	Гласные ([а], [о]), звонкие согласные ([м], [н], [в])
Неозвученные (unvoiced)	Турбулентный воздушный поток	Апериодический шум	Глухие согласные ([с], [ш], [п], [т]), шепот

На рисунке 1, *а*, *б* показана временная форма слова «Да» с фонетической сегментацией. Согласный [д], относящийся к озвученным (voiced) смычным согласным, характеризуется наличием регулярной голосовой периодичности: на временной форме видны колебания, связанные с вибрацией голосовых складок, а на спектрограмме (рисунок 1, *в, г*) прослеживаются гармонические полосы. Гласный [а], напротив, демонстрирует ярко выраженную периодическую структуру и формантные зоны, характерные для озвученных гласных. Таким образом, сопоставление согласного [д] и гласного [а] позволяет наглядно отразить различия акустических характеристик смычных и гласных звуков при общей принадлежности их к категории озвученных.

Эмоциональный стресс вызывает физиологические изменения в голосовом аппарате, дестабилизирующие гармоническую структуру речи:

– *повышенное напряжение голосовых складок* увеличивает F_0 и интенсивность сигнала [6, 9, 14];

– *учащенное и нестабильное дыхание* нарушает субглоттальное давление, повышая Jitter и Shimmer [1, 13, 15];

– *неполное смыкание складок* снижает гармоничность (HNR) и усиливает шум [2, 6];

– *напряжение артикуляторов* (язык, губы, челюсть) сужает резонаторы, изменяя форманты (например, $\uparrow F_1$, $\downarrow F_2$) [5, 14].

Эти изменения нарушают регулярность вибрации складок и фильтрацию резонаторов: гармоники ослабевают, шум возрастает, а формантные области искажаются [14-16]. Систематизация этих эффектов представлена в таблице 2, где приведены ключевые акустические параметры речи, чувствительные к эмоциональному стрессу, с указанием

физиологических причин их изменений, акустических проявлений и связанных дескрипторов.

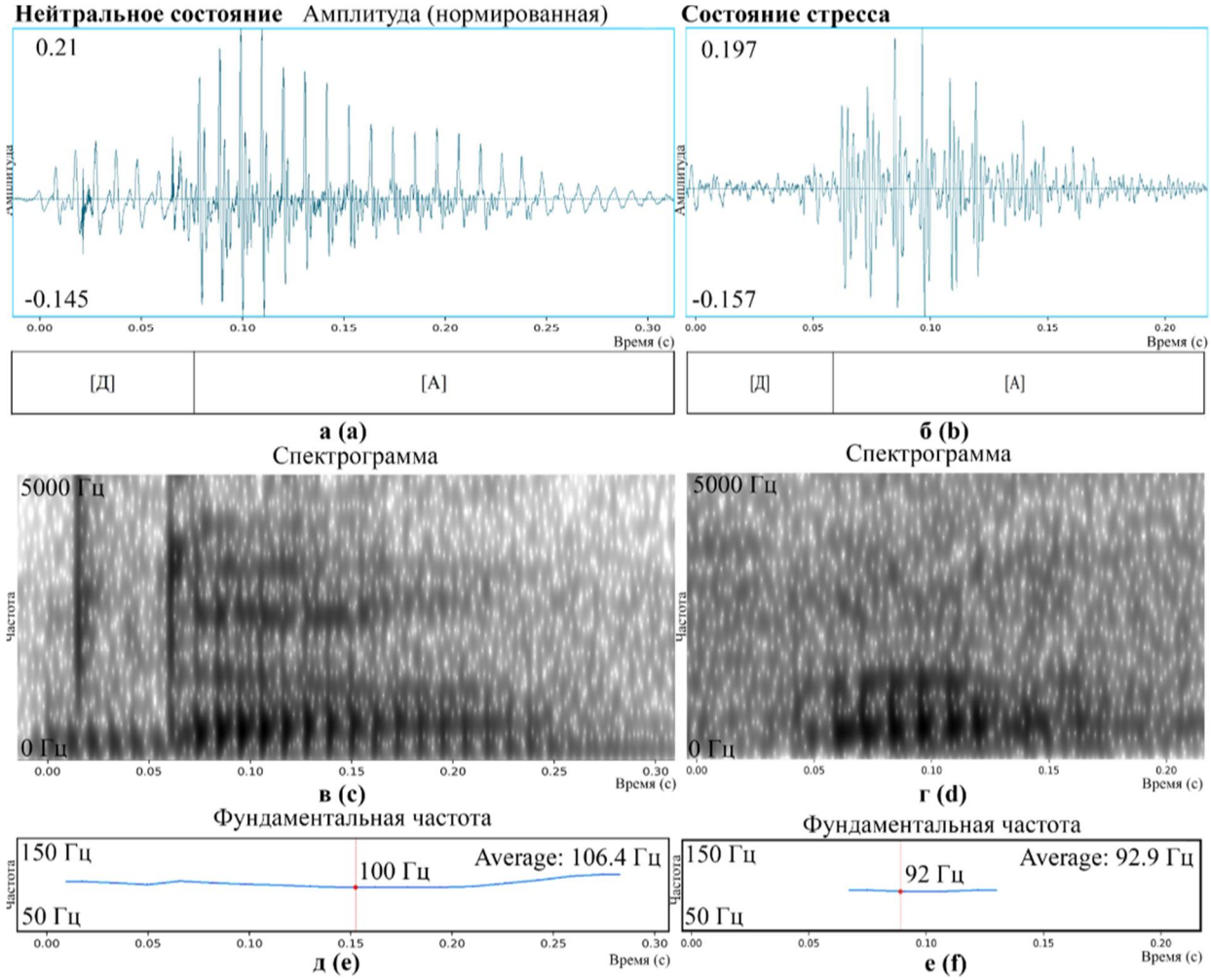

Рисунок 1 – Сравнение акустических характеристик слова «Да», произнесенного в нейтральном состоянии (слева, длительность 0,30 с) и в состоянии стресса (справа, длительность 0,21 с): а, б – временные формы сигнала с фонетической сегментацией; в, г – спектрограммы (по оси Y – частота, Гц); д, е – траектории фундаментальной частоты F_0 (Гц). По оси X на всех графиках – время (с) Анализ выполнен в программе Praat версия 6.4.34

Figure 1 – Comparison of acoustic characteristics of word «Yes» pronounced in a neutral state (left, duration 0,30 s) and in a stress state (right, duration 0,21 s): a, b – signal time shapes with phonetic segmentation; c, d – spectrograms (Y-axis – frequency, Hz); e, f – fundamental frequency trajectories F_0 (Hz). X-axis in all graphs is time (s). The analysis was performed in Praat version 6.4.34

Таблица 2 показывает, как стресс преобразуется в измеримые акустические метрики через физиологические реакции. Например, рост F_0 связан с напряжением голосовых складок [14], а снижение HNR – с их неполным смыканием [16], что усиливает шум. Jitter и Shimmer отражают нестабильность фонации [15], вызванную дыхательными и мышечными нарушениями. Методы анализа, такие как STFT и Cepstral Analysis, дополняют дескрипторы: STFT визуализирует динамику спектра [7], а Cepstral Analysis точно выделяет F_0 и форманты даже в сложных условиях [3]. Это делает акустические параметры надежным инструментом для оценки стресса.

Исследования подтверждают, что стресс повышает F_0 на 10 – 20 % и снижает гармоничность сигнала [13-15]. Дескрипторы (F_0, HNR, Jitter, Shimmer, Formant Analysis) обеспечивают количественную оценку этих изменений, что важно для клинической

диагностики и человеко-машинных систем, фиксируя влияние стресса на речь через объективные параметры голосообразования.

Таблица 2 – Влияние стресса на акустические параметры

Table 2 – Effect of stress on acoustic parameters

Дескрипторы	Физиологические причины при стрессе	Акустические проявления
F_0	Повышение тонуса крикоти-реоидной мышцы и субглотального давления при стрессе	Ускорение колебаний голосовых складок приводит к увеличению средней F_0 и ее вариабельности
σF_0	Нарушение нейромышечного контроля	Увеличение нестабильности F_0
HNR	Неполное смыкание голосовых складок и турбулентность воздуха при стрессе изменяют периодичность фонации	При «жесткой» напряженной фонации возрастает доля гармоник (HNR↑), при распаде смыкания – растет шумовая составляющая (HNR↓)
Jitter	Вариабельность мышечного тонуса и фонационных колебаний вызывает непостоянство длительности голосовых циклов	Проявляется в увеличении кратковременных колебаний частоты, растет относительный Jitter (процент пертурбаций частоты)
Shimmer	Флуктуации давления и артикуляторные изменения влияют на стабильность амплитуды фонации	Нестабильность амплитуды сопровождается ростом относительного Shimmer (процент пертурбаций амплитуды)
Formant Analysis	Изменение положения языка, губ и глотки при стрессе смещает резонансные частоты вокального тракта	Форма спектра меняется, сдвигаются частоты основных формант, могут меняться их ширина и расстояние между ними
STFT (spectrogram)	Отражает динамику спектра, дыхательная нестабильность и артикуляционная изменчивость при стрессе вносят быстрые изменения в спектр	Спектрограмма STFT выявляет смещение формант и усиление широкополосного шума при напряженной фонации
Cepstral Analysis	Чувствителен к изменениям периодичности источника звука и форме спектральной огибающей	Снижение кепстрального пика (CPP) указывает на усиление шумовой составляющей голоса из-за неполного смыкания голосовых складок, MFCC фиксируют форму спектральной огибающей

Экспериментальные исследования

Данные

В исследовании использовались две аудиозаписи голосового сигнала для графического сравнения параметров (рисунки 1, 2, 3). Обе записи были сделаны на устройстве Realmi C21. Испытуемый: мужчина, 39 лет. Параметры записи: WAV, 16-bit, 22 050 Hz.

Первая запись отражает *нейтральное состояние*, зафиксированное в контролируемой акустической среде без внешних раздражителей. Вторая запись относится к *стрессовому состоянию*, возникшему во время публичного выступления на судебном заседании. Данное состояние характеризовалось эмоциональной напряженностью, когнитивной нагрузкой и психологическим дискомфортом. Запись выполнена примерно через час после начала заседания, что позволяет интерпретировать его как *адаптивное состояние стресса*, когда организм уже частично приспособился к внешним условиям, сохраняя признаки повышенного напряжения [13].

Предобработка

Все аудиофайлы проходят единый конвейер препроцессинга:

– *нормализация громкости* – приведение RMS-уровня (среднеквадратической энергии) каждого файла к единому значению с помощью специального скрипта, написанного и исполненного в среде *PyCharm* [17];

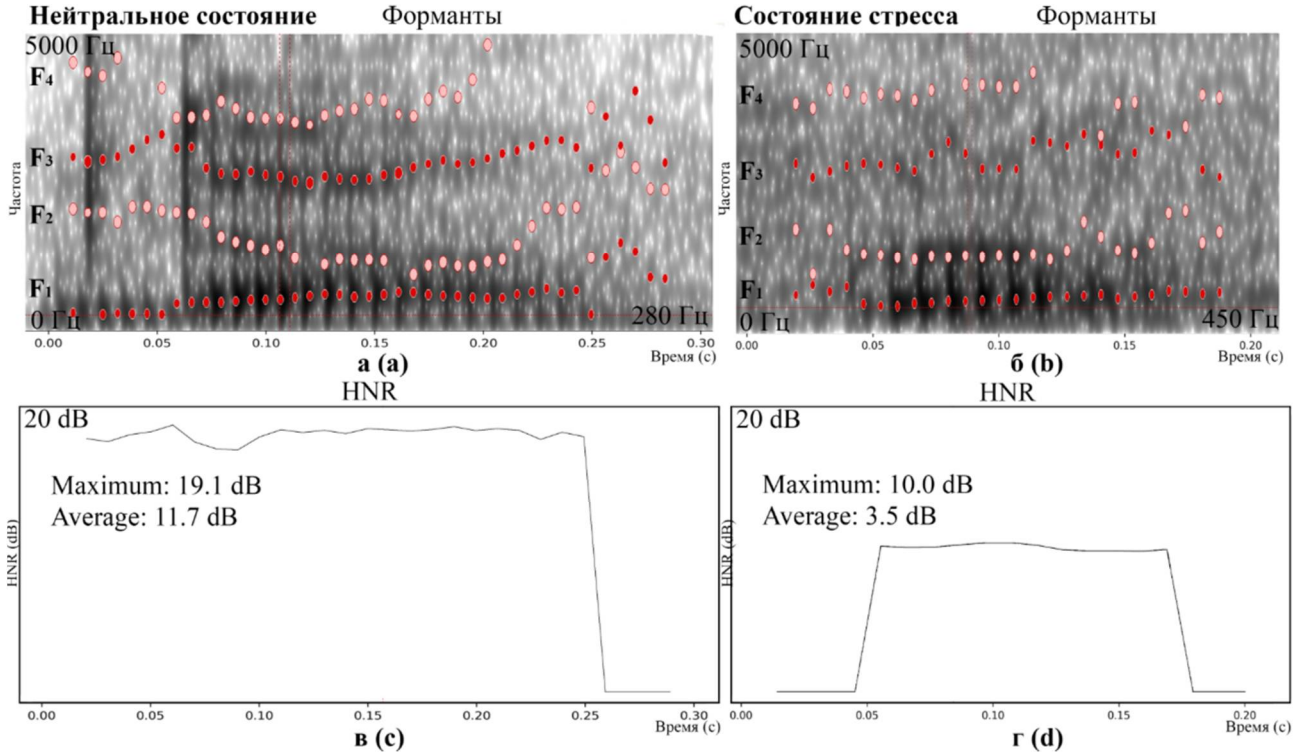

Рисунок 2 – Сравнение акустических характеристик слова «Да», произнесенного в нейтральном состоянии (слева, длительность 0,30 с) и в состоянии стресса (справа, длительность 0,21 с): а, б – формантные структуры F_1–F_4 (Гц); в, г – гармонико-шумовое отношение HNR (дБ). По оси X на всех графиках – время (с). Анализ выполнен в программе Praat версия 6.4.34

Figure 2 – Comparison of acoustic characteristics of the word «Yes» pronounced in a neutral state (left, duration 0,30 s) and under stress (right, duration 0,21 s): a, b – formant structures F_1–F_4 (Hz); c, d – harmonic-to-noise ratio HNR (dB). X-axis in all graphs is time (s). The analysis was performed in Praat version 6.4.34

– *очистка аудиосигнала* – для повышения качества дальнейшего анализа применялась процедура шумоподавления аудиосигналов [17].

Программное обеспечение

В ходе исследования использовались следующие программные средства и библиотеки:

– *iZotope RX 11* – специализированное программное обеспечение для предварительной обработки аудиозаписей, включая ресемплинг и выделение целевых фрагментов звукового сигнала;

– *Praat (версия 6.4.34)* – программный пакет для фонетического анализа речи, применявшийся для стандартизированного расчета акустических параметров и визуализации голосового сигнала [17];

– *PyCharm (версия 2023.2)* в связке с *Python (версия 3.9)* – интегрированная среда разработки, использовавшаяся для реализации собственных алгоритмов обработки речи, статистического анализа данных и построения графиков [17].

Библиотеки Python:

– *NumPy и Pandas* – для обработки и структурирования данных;

– *Matplotlib и Seaborn* – для построения графиков и визуализации результатов;

– *Noisereduce и Pydub* – для подавления шумов, нормализации громкости и предобработки аудиозаписей;

– *Parselmouth* – для анализа речевых сигналов с использованием функциональности Praat.

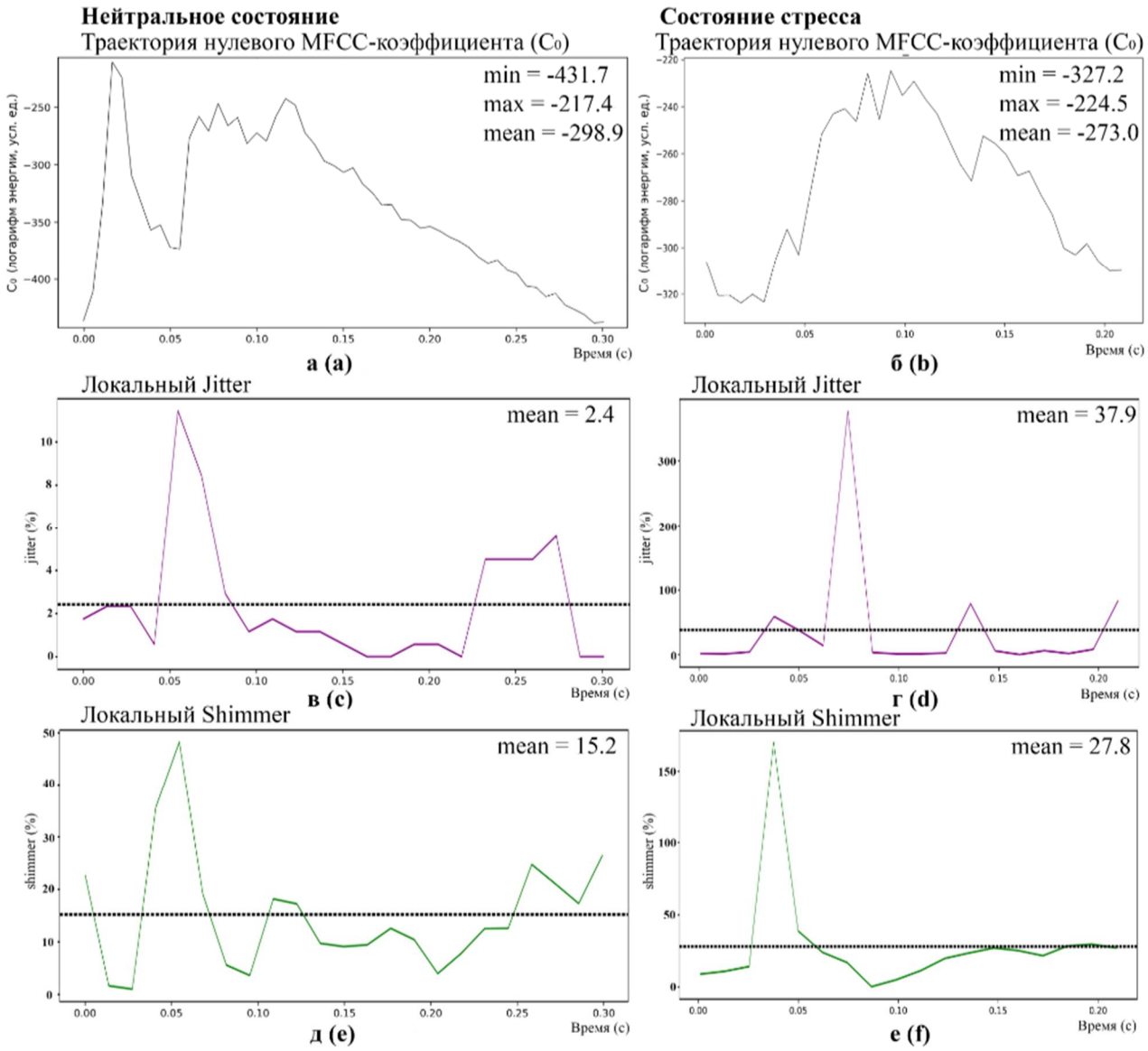

Рисунок 3 – Сравнение параметров голосового сигнала в нейтральном состоянии (слева, длительность 0,30 с) и в состоянии стресса (справа, длительность 0,21 с): а, б – траектории нулевого MFCC-коэффициента (C_0, логарифм энергии, усл. ед.); в, г – локальный Jitter (%), характеризующий относительные колебания длительности периода; д, е – локальный Shimmer (%), отражающий вариабельность амплитуды последовательных периодов. По оси X во всех графиках – время (с). Анализ выполнен в программе Python версия 3.9

Figure 3 – Comparison of voice signal parameters in neutral state (left, duration 0,30 s) and in stress state (right, duration 0,21 s): a, b – trajectories of zero MFCC coefficient (C_0, energy logarithm, arbitrary units); c, d – local jitter (%), characterizing relative fluctuations in period duration; e, f – local shimmer (%), reflecting the variability of amplitude of successive periods. X-axis in all graphs is time (s). The analysis was performed in Python version 3.9 version 3.9

Методы исследования

В работе проведен теоретический анализ акустических дескрипторов речи с учетом физиологии голосообразования, их алгоритмических основ и чувствительности к эмоцио-

нальным изменениям с использованием методов системного анализа и сравнительного обзора.

Экспериментальная часть включает сравнение двух аудиозаписей (нейтральной и стрессовой) одного испытуемого с использованием Praat и Python для извлечения и сопоставления ключевых параметров речи, с применением методов статистического анализа и визуализации данных.

Классические акустические дескрипторы

Основная частота (F_0)

Основная частота F_0 – это частота вибраций голосовых складок, определяющая высоту тона и являющаяся ключевым параметром фонации. Физиологически повышение F_0 и его вариабельности (σF_0) при эмоциональном возбуждении обусловлено увеличением тонуса крикотиреоидной мышцы и изменением субглоттального давления [13, 18]. F_0 надежно регистрирует стресс-индуцированные изменения и широко используется для детекции эмоциональных состояний [5, 19, 20].

Алгоритмы извлечения F_0 основаны на анализе периодичности сигнала. Наиболее распространенный метод – автокорреляция: определяется лаг (временной сдвиг) с максимальной корреляцией между сегментами сигнала, соответствующий среднему периоду вибрации, который затем инвертируется в частоту [3, 19, 21]. Альтернативные подходы включают кепстральный анализ, где F_0 определяется по положению первого пика в кепстральной области, а также методы гармонического поиска [6], использующие спектральное соответствие гармоник. Как правило, алгоритмы извлечения F_0 предполагают предварительную сегментацию сигнала и оконный анализ для обеспечения временной локализации. F_0 чувствителен к шумам и непериодическим участкам, но при корректной настройке обеспечивает высокую точность [8, 21].

На рисунке 1, *д, е* представлены траектории основной частоты (F_0) слова «Да» в нейтральном (рисунок 1, *д*) и стрессовом (рисунок 1, *е*) состояниях. По оси абсцисс отложено время (с), по оси ординат – F_0 (Гц). В нейтральном состоянии F_0 колеблется вблизи базового уровня (90 – 100 Гц) и демонстрирует плавное изменение во времени. В стрессовом состоянии наблюдается общее понижение F_0, что отражает повышение мышечного напряжения и возрастание тонуса голосовых связок.

Гармонико-шумовое отношение (HNR)

HNR (Harmonic-to-Noise Ratio) отражает степень регулярности фонации и соотношение энергии периодического (гармонического) и апериодического (шумового) компонентов речевого сигнала. Снижение HNR указывает на нестабильное смыкание голосовых складок и нарушение вибрационного паттерна, что характерно для состояний повышенного психофизиологического напряжения [2, 18, 22].

Методика расчета HNR может основываться на автокорреляционном анализе, где степень регулярности оценивается через максимумы корреляционной функции, либо на спектральной декомпозиции, при которой вычисляется отношение мощностей гармонической и шумовой составляющих в частотной области [8, 19]. Алгоритм HNR требует высокой точности определения границ фонации и устойчивого соотношения сигнал/шум. Этот параметр информативен при анализе нарушений вокальной стабильности, характерных для эмоционального возбуждения [3, 21].

На рисунке 2, *в, г* показаны значения гармонико-шумового отношения (HNR) для слова «Да» в нейтральном (рисунок 2, *в*) и стрессовом (рисунок 2, *г*) состояниях. По оси абсцисс отложено время (с), по оси ординат – HNR (дБ). В нейтральном состоянии HNR сохраняет высокие значения, что указывает на гармонически устойчивый сигнал с низкой шумовой компонентой. В стрессовом состоянии наблюдается снижение HNR, отражающее возрастание шумовой составляющей в голосе вследствие напряжения голосовых связок.

Джиттер (Jitter)

Jitter (дрожание частоты) характеризует микровариабельность длительности фонационных циклов, отражая нестабильность нейромышечного управления голосовыми складками. Повышенные значения Jitter фиксируются при нарушении центральной и периферической регуляции фонации, а также при дыхательной нестабильности, связанной со стрессом и активацией вегетативной нервной системы [3, 7, 18].

Алгоритм расчета Jitter включает детекцию границ фонационных циклов с последующим вычислением средней абсолютной разности длительностей соседних циклов, нормированной на средний период [8, 21]. Jitter измеряется в миллисекундах или процентах и высоко чувствителен к вариациям ритма фонации. Расчет требует высокой частоты дискретизации и точной предобработки сигнала, так как алгоритм расчета Jitter чувствителен к шуму и неточностям детекции циклов [5]. Вычислительная сложность Jitter умеренная, но зависит от качества сегментации сигнала [4, 5, 19].

На рисунке 3, на фрагментах *в, г* показаны изменения локального Jitter (%). В стрессовом состоянии среднее значение Jitter, равное 37,9 %, значительно выше по сравнению с нейтральным состоянием (2,4 %), что указывает на рост нерегулярности вибрации голосовых складок под влиянием стресса.

Шиммер (Shimmer)

Shimmer (дрожание амплитуды) оценивает микровариабельность амплитуды фонационных импульсов и отражает нестабильность вокального усилия. Амплитудные колебания, усиливающиеся при стрессовых состояниях, свидетельствуют о нарушениях дыхательно-фонационного взаимодействия и нейромышечной координации [7, 18, 22]. Таким образом, Shimmer является индикатором стабильности вокального усилия.

Алгоритм расчета Shimmer основан на анализе амплитудной последовательности: определяется средняя абсолютная разность амплитуд между соседними фонационными импульсами, нормированная на среднюю амплитуду сигнала [3, 21]. Для достоверного извлечения параметра требуются амплитудная нормализация, высокая точность в определении фонационных границ и фильтрация фонового шума. Повышенный Shimmer может указывать на снижение фонационной устойчивости и увеличение вариативности акустического давления [8, 19, 20].

На рисунке 3, на фрагментах *д*, *е* отображены локальные значения Shimmer (%) в нейтральном и стрессовом состояниях. В стрессовом состоянии среднее значение Shimmer, равное 27,8 %, выше, чем в нейтральном состоянии (15,2 %), что свидетельствует о большей амплитудной нестабильности сигнала при стрессовом произнесении.

Формантный анализ (Formant Analysis)

Форманты (F_1, F_2, F_3 и др.) представляют собой частоты резонансных максимумов голосового тракта, определяемые его артикуляционной конфигурацией [5, 13, 18]. Первые три форманты (F_1, F_2, F_3) несут основную речевую информацию, так как именно они обеспечивают различие гласных и значительную часть акустических характеристик согласных. Эмоциональное возбуждение сопровождается изменениями в положении языка, глотки и ротовой полости, что вызывает смещение формантных частот [19, 23].

Анализ формантов реализуется через построение спектральной огибающей сигнала. Основной метод – линейное предсказание (LPC) [11], при котором сигнал моделируется авторегрессионной системой, а ее полюса интерпретируются как формантные частоты. Альтернативно применяется прямой спектральный анализ с применением быстрого преобразования Фурье (FFT), где форманты извлекаются как локальные максимумы спектра [5]. Оба описанных подхода требуют аккуратной сегментации целевых фрагментов речи и достаточной частоты дискретизации. Анализ динамики формант позволяет косвенно оценивать уровень эмоционального напряжения через выявление артикуляторной нестабильности [21].

На рисунке 2, *а, б* представлены траектории формант (F_1-F_4) слова «Да» в нейтральном (*а*) и стрессовом (*б*) состояниях. По оси абсцисс отложено время (с), по оси ординат – частота формант (Гц). В нейтральном состоянии формантные траектории более стабильны, тогда как в стрессовом состоянии наблюдаются колебания и смещения частотных максимумов, особенно для F_1 и F_2.

Кепстральный анализ (Cepstral Analysis)

Кепстральный анализ позволяет декомпозировать сигнал на быстро изменяющиеся компоненты глоттального возбуждения (включая F_0) и медленно изменяющиеся компоненты спектральной огибающей, отражающие формантную структуру голосового тракта [18]. Метод широко применяется для оценки стабильности фонации в условиях нестабильности голоса [19].

Метод основан на преобразовании логарифма амплитудного спектра сигнала с последующим обратным преобразованием Фурье, что позволяет представить сигнал в кепстральной области [10]. Ключевыми параметрами являются *Cepstral Peak Prominence (CPP)*, характеризующий четкость и амплитуду основного пика, связанного с регулярностью фонации и периодической структуры сигнала, и *Mel Frequency Cepstral Coefficients (MFCC)*, описывающие форму спектра в перцептивной шкале Mel, приближенной к особенностям слухового восприятия человека [14]. Снижение CPP указывает на нарушение регулярности фонации, тогда как MFCC чувствительны к артикуляционным изменениям, связанным с эмоциональными состояниями [3]. Кепстральные методы отличаются устойчивостью к шумам [8] и высокой чувствительностью при анализе спектральной динамики речи.

На рисунке 3, на фрагментах *а*, *б* представлены траектории нулевого MFCC-коэффициента (C_0), интерпретируемого как логарифм энергии сигнала. В стрессовом состоянии (рисунок 3, *б*) наблюдаются более высокие значения C_0 по сравнению с нейтральным состоянием (рисунок 3, *а*), что отражает увеличение энергии и интенсивности звучания.

Спектральный анализ на основе STFT

Коротковременное преобразование Фурье (STFT) применяется для исследования временной эволюции спектра речевого сигнала. Метод основан на разбиении сигнала на короткие перекрывающиеся окна длительностью 20 – 40 мс с коэффициентом перекрытия 50 – 75 % [9], для каждого из которых вычисляется спектр. Результатом является спектрограмма – двумерное представление, отображающее изменение амплитуд частотных компонентов во времени [19, 21]. Разрешение STFT зависит от размера окна, что требует баланса между временной и частотной точностью.

STFT обеспечивает информативную визуализацию динамики спектра, включая усиление шумовой составляющей, снижение выраженности гармоник и смещение формантных частот, что характерно для состояний эмоционального напряжения [5, 19, 23]. Несмотря на то, что метод не извлекает специфических акустических дескрипторов напрямую, он используется как вспомогательный инструмент для качественного анализа временно-частотных характеристик речи в акустической диагностике и исследовании психоэмоциональных состояний [5].

На рисунке 1, *в*, *г* представлены спектрограммы слова «Да» в нейтральном (*в*) и стрессовом (*г*) состояниях. По оси абсцисс отложено время (с), по оси ординат – частота (Гц). Более высокая концентрация энергии во времени и усиление высокочастотных компонентов наблюдаются в стрессовом состоянии по сравнению с нейтральным.

Классические акустические дескрипторы (F_0, HNR, Jitter, Shimmer, Formant Analysis, Cepstral Analysis, STFT) обеспечивают многопараметрический анализ гармонической структуры речи, охватывая источник звука, фильтрационные и спектральные свойства [5, 10, 19]. Они взаимодополняют друг друга, выявляя стресс-индуцированные изменения фонации и артикуляции [8]. Благодаря физиологической интерпретируемости и верифицируемости результатов, эти параметры формируют основу для разработки объяснимых систем

диагностики эмоционального стресса, превосходящих черный ящик нейросетевых моделей в клинически значимых приложениях [5].

В таблице 3 показаны основные наблюдаемые эффекты при переходе от нейтрального к стрессовому произнесению в исследуемом примере: существенное падение гармоничности (HNR), сильный рост микровариабельностей (Jitter, Shimmer) и увеличение общей энергичности сигнала ($MFCC_0$). Средняя F_0 в этом конкретном примере уменьшилась, что подчеркивает индивидуальную вариативность реакций голоса на стресс и необходимость осторожной интерпретации результатов, полученных на ограниченном материале.

Таблица 3 – Сводные результаты сравнения акустических параметров

Table 3 – Summary results of comparison of acoustic parameters

Показатель	**Единицы**	**Нейтральное состояние**	**Состояние стресса**	**Направление изменения**
F_0 (средняя)	Гц	106,4	92,9	снижение у данного испытуемого
HNR (среднее)	дБ	11,7	3,5	снижение гармоничности
Jitter (local)	%	2,4	37,9	резкое возрастание нестабильности частоты
Shimmer (local)	%	15,2	27,8	рост амплитудной нестабильности
$MFCC_0$ (C_0, log-energy)	усл. ед.	−298,9	−273,0	увеличение энергии
Форманты (динамика)	Гц	стабильные траектории F_1-F_4	увеличенная вариабельность, смещения ($\uparrow F_1$, $\downarrow F_2$)	возрастание вариабельности

Таблица 3 показывает основные наблюдаемые эффекты при переходе от нейтрального к стрессовому состоянию в исследуемом примере у одного испытуемого – существенное падение гармоничности (HNR), сильный рост микровариабельностей (Jitter, Shimmer) и увеличение общей энергичности сигнала ($MFCC_0$). Средняя F_0 в этом конкретном примере уменьшилась, что подчеркивает индивидуальную вариативность реакций голоса на стресс и необходимость осторожной интерпретации результатов, полученных на ограниченном материале.

Полученные в данном исследовании изменения акустических параметров носят иллюстративный характер в силу ограниченного объема выборки и согласуются с результатами, полученными на более репрезентативных данных в других работах. Так, в работах [13, 18] подтверждается, что эмоциональный стресс статистически значимо приводит к увеличению вариабельности основного тона (σF_0) и параметров микровариабельностей (Jitter, Shimmer). Снижение гармоничности сигнала (HNR) при стрессе, связанное с нарушением регулярности смыкания голосовых складок, также отмечается в исследованиях [14, 18]. Повышенная вариабельность формантных частот (F_1 – F_4), отражающая артикуляционную нестабильность, и рост энергии сигнала, фиксируемый нулевым коэффициентом $MFCC_0$, также являются установленными акустическими коррелятами стрессового состояния [7, 14, 18]. Таким образом, наблюдаемые в данном эксперименте направления изменений по всему спектру дескрипторов – от параметров источника до спектральных характеристик – соответствуют установленным в литературе эффектам.

Таким образом, полученные результаты формируют представление об изменениях голоса под воздействием стресса. Зафиксированное снижение частоты основного тона можно объяснить состоянием контролируемого стресса: от начала стрессовой ситуации до момента аудиозаписи прошло около одного часа.

Заключение

В настоящей работе систематизированы классические акустические дескрипторы гармонической структуры речи – F_0, HNR, Jitter, Shimmer, форманты (Formant Analysis), STFT и кепстральные коэффициенты (Cepstral Analysis, MFCC*),* рассмотрены их физиологические основания, алгоритмические подходы к извлечению и чувствительность к эмоциональному стрессу. На демонстрационных примерах (один испытуемый, две записи) выявлены типичные для стрессовой фонографии изменения: снижение периодичности фонации (снижение HNR), заметное повышение микровариабельностей частоты и амплитуды (повышение Jitter и Shimmer), а также возрастание энергии сигнала (повышение $MFCC_0$). Динамика формант показала артикуляторную нестабильность (увеличение разброса и локальные смещения $F_1 – F_4$). Факт понижения средней F_0 в данном примере согласуется с интерпретацией состояния контролируемого стресса (запись сделана примерно через час после начала стрессора) и указывает на важность учета временной динамики реакций.

Практические наблюдения показывают, что комбинированный подход – совместное использование параметров источника, отражающих состояние и динамику голосовых складок (F_0, Jitter, Shimmer), показателей гармоничности (HNR) и спектрально-кепстральных характеристик (MFCC, CPP, Formant), позволяет получить более полное и физиологически интерпретируемое описание голосовой реакции на стресс по сравнению с использованием одиночных дескрипторов.

Библиографический список

1. **Schuller B., Batliner A., Bergler C.** et al. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks // Proc. Interspeech 2020. 2020. Pp. 2042-2046. DOI: 10.21437/Interspeech.2020-32.

2. **Huang Z., Dong M., Mao Q., Zhan Y.** Speech Emotion Recognition Using CNN // Proceedins of the 22nd ACM international conference on Multimedia. 2014. Pp. 801-804. DOI: 10.1145/2647868.2654984.

3. **Zhao J., Mao X., Chen L.** Speech emotion recognition using deep 1D & 2D CNN LSTM networks // Biomedical Signal Processing and Control. 2019. Vol. 47. № 4. Pp. 312-323. DOI: 10.1016/ j.bspc.2018.08.035.

4. **Cummins N., Scherer S., Krajewski J.**et al. A review of depression and suicide risk assessment using speech analysis // Speech Communication. 2015. Vol. 71. Pp. 10-49. DOI: 10.1016/j.specom.2015.03.004.

5. **Tzirakis P., Trigeorgis G., Nicolaou M.A.** et al. End-to-end speech emotion recognition using deep neural networks // IEEE Journal of Selected Topics in Signal Processing. 2017. Vol. 11. № 8. Pp. 1301-1309. DOI: 10.1109/JSTSP.2017.2764438.

6. **Drugman T., Alwan A.** Joint robust voicing detection and pitch estimation based on residual harmonics // Proceedings of Interspeech. 2011. Pp. 1973-976. DOI: 10.21437/Interspeech.2011-519.

7. **Ververidis D., Kotropoulos C.** Emotional speech recognition: Resources, features, and methods // Speech Communication. 2006. Vol. 48. № 9. Pp. 1162-1181. DOI: 10.1016/j.specom.2006.04.003.

8. **Eyben F., Scherer K.R., Schuller B.** et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing // IEEE Transactions on Affective Computing. 2015. Vol. 7. № 2. Pp. 190-202. DOI: 10.1109/TAFFC.2015.2457417.

9. **Kwon O.-W., Chan K., Hao J., Lee T.-W.** Emotion recognition by speech signals // Proc. 8th European Conference on Speech Communication and Technology (ICASSP). 2003. Pp. 125-128. DOI: 10.21437/Eurospeech.2003-80.

10. **Алимурадов А.К., Чураков П.П.** Обзор и классификация методов обработки речевых сигналов в системах распознавания речи // Измерение. Мониторинг. Управление. Контроль. 2015. № 2 (12). С. 27-35. URL: https://cyberleninka.ru/article/n/obzor-i-klassifikatsiya-metodov-obrabotki-rechevyh-signalov-v-sistemah-raspoznavaniya-rechi) (дата обращения: 20.08.2025).

11. **Lehoux H.** Biomechanics and acoustics of voice production: doctoral dissertation / H. Lehoux; Study program «Biophysics», supervisor J. G. Svec. 2023. 160 p.

12. **Koreman J.J.M.** Decoding linguistic information in the glottal airflow: doctoral dissertation. Leiden, 1996. 120 p.

13. **Лебедева Н.Н., Каримова Е.Д.** Акустические характеристики речевого сигнала как показатель функционального состояния человека // Успехи физиологических наук. 2014. Т. 45. № 1. С. 57-95.

14. **Van Puyvelde M., Neyt X., McGlone F.** et al. Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance [Electronic resource] // Frontiers in Psychology. 2018. Vol. 9, Art. 1994. DOI: 10.3389/fpsyg.2018.01994.

15. **Wang Q., Xu F., Wang X.** et al. How Anxiety State Influences Speech Parameters: A Network Analysis Study from a Real Stressed Scenario [Electronic resource] // Behavioral es. 2025. Vol. 15. № 3. P. 262. DOI: 10.3390/brainsci15030262.

16. **Titze I.R.** Workshop on Acoustic Voice Analysis: Summary Statement [Electronic resource]. Denver: National Center for Voice and Speech, 1995.

17. **Jadoul Y., Thompson B., de Boer B.** Introducing Parselmouth: A Python interface to Praat // Journal of Phonetics. 2018. Vol. 7. Pp. 1-15. DOI: 10.1016/j.wocn.2018.07.001.

18. **Schewski L., Magimai-Doss M., Beldi G., Keller S.** Measuring negative emotions and stress through acoustic correlates in speech: a systematic review // PLoS ONE. 2025. Vol. 20. № 7. Art. e0328833. DOI: 10.1371/journal.pone.0328833.

19. **Koffi E.** A comprehensive review of jitter, shimmer, and HNR: Linguistic and paralinguistic applications // Linguistic Portfolios. 2025. Vol. 14, Article 2. URL: https:// repository.stcloudstate.edu/ling/vol14/iss1/2 (дата обращения: 27.08.2025).

20. **Stas J., Ondas S., Juhar J.** Performance evaluation of different speech-based emotional stress level detection approaches // IEEE Access. 2025. DOI: 10.1109/ACCESS.2025.3584534.

21. **Singh P., Sahidullah M. Saha G.** Modulation spectral features for speech emotion recognition using deep neural networks // Speech Communication. 2023. Vol. 146. № 16, Pp. 53-69. DOI: 10.1016/j.specom.2022.10.003.

22. **Субботина М.В., Заббарова И.Б.** Влияние длины голосовых складок на диапазон голоса у начинающих вокалистов // Журнал прикладной акустики. 2018. № 2. С. 23-30. URL: https:// cyberleninka.ru/article/n/vliyanie-dliny-golosovyh-skladok-na-diapazon-golosa-u-nachinayu-schih-vokalistov (дата обращения: 15.08.2025).

23. **Banse R., Scherer K.R.** Acoustic profiles in vocal emotion expression // Journal of Personality and Social Psychology. 1996. Vol. 70. No. 3. Pp. 614-636. DOI: 10.1037/0022-3514.70.3.614.

UDC 004.934:681.518

ACOUSTIC DESCRIPTORS OF HARMONIC SPEECH STRUCTURE FOR EMOTION ASSESSMENT

O. V. Melnik, Dr. in technical sciences, full professor, RSREU, Ryazan, Russia;
orcid.org/0000-0002-3513-2180, e-mail: omela111@yandex.ru
S. I. Babaev, Ph. D. (technical sciences.), associate professor, RSREU, Ryazan, Russia;
orcid.org/0000-0001-5829-8223, e-mail: babaev.s.i@gmail.com
M. N. Saraev, post-graduate student, RSREU, Ryazan, Russia;
orcid.org/0009-0006-5118-3478, e-mail: mixailr@mail.ru

The article presents classical acoustic descriptors based on the harmonic structure of speech used for automatic assessment of emotional states (neutral state – stress). The aim of the work is to systematize methods for analyzing the harmonic structure of speech, reveal their physiological basis and assess their informativeness with respect to emotional changes. Key methods are considered: analysis of harmonic-to-noise ratio (HNR), estimation of fundamental frequency (F_0), parameters characterizing instability of period and amplitude (Jitter and Shimmer), spectral analysis based on short-time Fourier transform (STFT), cepstral analysis and formant analysis. Their extraction algorithms and sensitivity to emotional changes are described. Particular emphasis is placed on physiologically interpretable parameters (F_0, HNR, Jitter and Shimmer) and on the fundamental methods underlying their calculation - spectral and cepstral analysis. The limitations of each method are highlighted, and recommendations for selecting descriptors are provided. The practical significance of the methods discussed lies in demonstrating their applicability on illustrative mate-

rial: in a paired comparison (neutral - stress) characteristic changes were observed - a decrease in HNR, increases in Jitter and Shimmer, an increase in signal energy ($MFCC_0$), and greater formant variability ($F_1 – F_4$). This confirms the sensitivity of the descriptors to emotional stress and supports the use of a combined feature set. The article will be useful to specialists in signal processing, psycholinguistics and emotion recognition systems.

***Keywords**: speech harmonics, HNR method, Fundamental Frequency estimation, Jitter, Shimmer, STFT analysis, Formant Analysis, Cepstral Analysis, emotion assessment, stress.*

DOI: 10.21667/1995-4565-2026-95-171-185

References

1. **Schuller B., Batliner A., Bergler C.** et al. The INTERSPEECH 2020 Computational Paralinguistics Challenge: *Elderly Emotion, Breathing & Masks. Proc. Interspeech 2020*, pp. 2042-2046. DOI: 10.21437/Interspeech.2020-32.

2. **Huang Z., Dong M., Mao Q., Zhan Y.** Speech Emotion Recognition Using CNN. *Proceedins of the 22nd ACM international conference on Multimedia*. 2014, pp. 801-804. DOI: 10.1145/2647868.2654984.

3. **Zhao J., Mao X., Chen L.** Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*. 2019, vol. 47, no. 4, pp. 312-323. DOI: 10.1016/ j.bspc.2018.08.035.

4. **Cummins N., Scherer S., Krajewski J.**et al. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*. 2015, vol. 71, pp. 10-49. DOI: 10.1016/ j.specom.2015.03.004.

5. **Tzirakis P., Trigeorgis G., Nicolaou M.A.** et al. End-to-end speech emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*. 2017, vol. 11, no. 8, pp. 1301-1309. DOI: 10.1109/JSTSP.2017.2764438.

6. **Drugman T., Alwan A.** Joint robust voicing detection and pitch estimation based on residual harmonics. *Proceedings of Interspeech*. 2011, pp. 1973-976. DOI: 10.21437/Interspeech.2011-519.

7. **Ververidis D., Kotropoulos C.** Emotional speech recognition: Resources, features, and methods. *Speech Communication*. 2006, vol. 48, no. 9, pp. 1162-1181. DOI: 10.1016/j.specom.2006.04.003.

8. **Eyben F., Scherer K.R., Schuller B.** et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*. 2015, vol. 7, no. 2, pp. 190-202. DOI: 10.1109/TAFFC.2015.2457417.

9. **Kwon O.-W., Chan K., Hao J., Lee T.-W.** Emotion recognition by speech signals. *Proc. 8th European Conference on Speech Communication and Technology (ICASSP)*. 2003, pp. 125-128. DOI: 10.21437/Eurospeech.2003-80.

10. **Alimuradov A.K., Churakov P.P.** Obzor i klassifikaciya metodov obrabotki rechevykh signalov v sistemakh raspoznavaniya rechi. *Izmerenie. Monitoring. Upravlenie. Kontrol'*. 2015, no. 2 (12), pp. 27-35. URL: https://cyberleninka.ru/article/n/obzor-i-klassifikatsiya-metodov-obrabotki-rechevyh-signalov-v-sistemah-raspoznavaniya-rechi) (data obrashcheniya: 20.08.2025).

11. **Lehoux H.** Biomechanics and acoustics of voice production: doctoral dissertation; *Study program «Biophysics», supervisor J. G. Svec*. 2023. 160 p.

12. **Koreman J.J.M.** Decoding linguistic information in the glottal airflow: doctoral dissertation. Leiden, 1996. 120 p.

13. **Lebedeva N.N., Karimova E.D.** Akusticheskie kharakteristiki rechevogo signala kak pokaza-tel' funkcional'nogo sostoyaniya cheloveka. *Uspekhi fiziologicheskikh nauk.* 2014, vol. 45, no. 1, pp. 57-95.

14. **Van Puyvelde M., Neyt X., McGlone F.** et al. Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance [*Electronic resource*]. *Frontiers in Psychology*. 2018, vol. 9. Art. 1994. DOI: 10.3389/fpsyg.2018.01994.

15. **Wang Q., Xu F., Wang X.** et al. How Anxiety State Influences Speech Parameters: A Network Analysis Study from a Real Stressed Scenario [*Electronic resource*]. *Behavioral Sciences*. 2025, vol. 15, no. 3, p. 262. DOI: 10.3390/brainsci15030262.

16. **Titze I.R.** Workshop on Acoustic Voice Analysis: Summary Statement [*Electronic resource]. Denver: National Center for Voice and Speech.* 1995.

17. **Jadoul Y., Thompson B., de Boer B.** Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*. 2018, vol. 7, pp. 1-15. DOI: 10.1016/j.wocn.2018.07.001.

18. **Schewski L., Magimai-Doss M., Beldi G., Keller S.** Measuring negative emotions and stress through acoustic correlates in speech: a systematic review. *PLoS ONE*. 2025, vol. 20, no. 7. Art. e0328833. DOI: 10.1371/journal.pone.0328833.

19. **Koffi E.** A comprehensive review of jitter, shimmer, and HNR: Linguistic and paralinguistic applications. *Linguistic Portfolios*. 2025, vol. 14. Article 2. URL: https://repository.stcloudstate.edu/ling/vol14/iss1/2 (дата обращения: 27.08.2025).

20. **Stas J., Ondas S., Juhar J.** Performance evaluation of different speech-based emotional stress level detection approaches. *IEEE Access*. 2025. DOI: 10.1109/ACCESS.2025.3584534.

21. **Singh P., Sahidullah M., Saha G.** Modulation spectral features for speech emotion recognition using deep neural networks. *Speech Communication.* 2023, vol. 146, no. 16, pp. 53-69. DOI: 10.1016/j.specom.2022.10.003.

22. **Subbotina M.V., Zabbarova I.B.** Vliyanie dliny golosovykh skladok na diapazon golosa u nachinayushchikh vokalistov. *Zhurnal prikladnoj akustiki*. 2018, no. 2, pp. 23-30. URL: https://cyberleninka.ru/article/n/vliyanie-dliny-golosovyh-skladok-na-diapazon-golosa-u-nachinayuschih-vokalistov (data obrashcheniya: 15.08.2025).

23. **Banse R., Scherer K.R.** Acoustic profiles in vocal emotion expression . *Journal of Personality and Social Psychology*. 1996, vol. 70, no. 3, pp. 614-636. DOI: 10.1037/0022-3514.70.3.614.