Обзорная статья
УДК 81.322
DOI: 10.17072/1993-0550-2025-2-101-122
https://elibrary.ru/ykakqn

Анализ подходов к автоматизации разметки паралингвистических характеристик в русскоязычных речевых данных

Евгений Николаевич Радченко[1], Екатерина Владимировна Исаева [2]
[1]Национальный исследовательский технологический университет "МИСИС", г. Москва, Россия
turnipseason@gmail.com
[2]Пермский государственный национальный исследовательский университет, г. Пермь, Россия
ekaterinaisae@psu.ru

Аннотация. Разработка систем синтеза речи с возможностью управления речевыми характеристиками посредством естественного языка имеет практический интерес, поскольку предоставляет интуитивно понятный способ влияния на результат генерации. Вместе с тем, для русскоязычных данных наблюдается недостаток как подобных систем, так и размеченных наборов данных, необходимых для их создания. Ручная разметка больших наборов данных является ресурсоемким процессом, требующим не только экспертных знаний предметной области, но и согласованности разметчиков между собой. В связи с этим, актуальным является исследование подходов к автоматизации аннотации паралингвистических характеристик русскоязычной речи, позволяющих унифицировать существующую разметку и ускорить ее масштабирование. В данной статье рассмотрены основные подходы к разметке таких паралингвистических характеристик, как паузы, ударения, а также высота и тембр голоса. Особое внимание уделено обзору доступных программных реализаций описанных методов. Ключевым выводом по итогам анализа стало наличие достаточного количества программных средств, пригодных для аннотации "базовых" характеристик в русскоязычной речи. Паузы и фундаментальная частота могут выделяться с помощью методов, не использующих лингвистическую информацию, в то время как для разметки ударений существуют методы, основанные на нейронных сетях и учитывающие контекст высказывания для снятия омографии, достигающие значения метрики Accuracy в 98%. В то же время автоматическая разметка более сложных характеристик, таких как тембр и выражаемые эмоции, остается малоизученной. Данные результаты указывают на необходимость дополнительных исследований в области методов автоматической аннотации паралингвистических характеристик в русскоязычных речевых данных.

Ключевые слова: *автоматическая аннотация; разметка аудио; разметка текста; паралингвистические характеристики; генерация речи*

Для цитирования: *Радченко Е. Н., Исаева Е. В*. Анализ подходов к автоматизации разметки паралингвистических характеристик в русскоязычных речевых данных // Вестник Пермского университета. Математика. Механика. Информатика. 2025. № 2(69). С. 101−122. DOI: 10.17072/1993-0550-2025-2-101-122. https://elibrary.ru/ykakqn

Статья поступила в редакцию 28.04.2025; одобрена после рецензирования 15.06.2025; принята к публикации 11.07.2025.

Review article

Analysis of Approaches to Paralinguistic Feature Annotation Automation in Russian Speech

Evgenii. N. Radchenko[1], Ekaterina. V. Isaeva[2]
[1]National University of Science and Technology "MISIS", Moscow, Russia
urnipseason@gmail.com
[2]Perm State University, Perm, Russia
ekaterinaisae@psu.ru

Abstract. The development of speech synthesis systems with the ability to control speech characteristics using natural language is of practical interest, since it provides an intuitive way to influence the results of the generation. At the same time, for Russian-language data there exists a shortage of both such systems and labeled datasets required to create them. Manual labeling of large datasets is a resource-intensive process that requires not only expert knowledge, but also inter-annotator labeling consistency. In this regard, the task of automating the annotation of paralinguistic characteristics of Russian-language speech becomes relevant, allowing to unify the labeling already existing in available datasets as well as accelerate its scaling to unlabeled ones.

This article considers the main approaches to the annotation of such paralinguistic characteristics as pauses, stresses, as well as the pitch and timbre of the voice. In particular, attention is paid to reviewing available software implementations of the methods described.

The key conclusion from the analysis was the existence of a sufficient number of methods suitable for annotating "basic" characteristics in Russian-language speech. Pauses and fundamental frequency can be extracted using methods that do not use linguistic information, while for stress annotation there are methods based on neural networks and, thus, taking into account the context of the utterance to resolve stress placement in homographs, achieving an Accuracy metric score as high as 98%. At the same time, automatic annotation of more complex characteristics, such as timbre and expressed emotions, remains poorly studied. These results indicate the need for additional research in the field of methods for automatic annotation of paralinguistic features in Russian-language speech corpora.

Keywords: *automatic annotation; audio annotation; text annotation; paralinguistic characteristics; speech generation*

For citation: Radchenko, E. N. and Isaeva, E. V. (2025), "Analysis of Approaches to Paralinguistic Feature Annotation Automation in Russian Speech", *Bulletin of Perm University. Mathematics. Mechanics. Computer Science*, no. 2(69), pp. 101−122. (In Russ.). DOI: 10.17072/1993-0550-2025-2-101-122. https://elibrary.ru/ EDN

The article was submitted 28.04.2024; approved after reviewing 15.06.2024; accepted for publication 11.07.2025.

1. Введение

Естественная речь характеризуется спектром так называемых паралингвистических характеристик: длительностью пауз, ударениями, тембром и выражаемыми эмоциями говорящего, а также другими особенностями. В англоязычной литературе совокупность этих характеристик также называется *речевым стилем*. Современные алгоритмы генерации речи, основанные на глубоких нейросетях, такие как ParlerTTS и LibriTTSP [1, 2, 3], позволяют задавать желаемые характеристики синтезированного голоса, используя короткие текстовые описания.

Для использования такого способа задания стиля необходимы размеченные аудиотекстовые корпуса, включающие транскрипции и аннотацию соответствующих признаков. В то время как такие корпуса активно разрабатываются для англоязычных данных, существующие русскоязычные корпуса, такие как Dusha, ПРуД, РИНКО [4, 5, 6], обычно разработаны под конкретные узконаправленные задачи, и требуют дополнительной разметки для полноценного использования вышеобозначенных методов.

Таким образом, исследование подходов к автоматизации стилевой разметки представляется перспективным направлением, способным существенно упростить создание обучающих данных. Данная статья направлена на анализ существующих подходов к автоматической разметке паралингвистических характеристик речи и оценку их применимости к русскоязычным данным. Поскольку понятие паралингвистических характеристик охватывает широкий спектр различных качеств голоса, материал статьи поделен на подразделы, каждый из которых анализирует методы автоматизации разметки применительно к конкретным характеристикам. Для удобства чтения названия метрик оценки качества работы моделей машинного обучения в статье приводятся на английском.

2. Методы решения задачи автоматизации разметки паралингвистических характеристик

2.1. Автоматическая разметка пауз

В данном разделе будут рассмотрены методы моделирования невокализованных пауз (синтаксических и хезитационных пауз [7], если они не являлись заполненными).

Наиболее прямолинейным подходом к решению данной задачи является выделение синтаксических пауз, исходя из пунктуации. Стоит отметить, что в существующих наборах данных транскрибация зачастую получена путем автоматического распознавания речи. Так, например, в расшифровках аудио в датасете Dusha, пунктуация отсутствует. Для восстановления разметки могут быть использованы существующие программные решения, такие как "Восстановление пунктуации для русского языка" [8] и "Punctuation and casing restoration for the Russian Language (BERT-based)" [9] на основе нейронных сетей.

Наилучшее качество по метрике F1 (среднего гармонического между метриками Precision и Recall) такие модели показывают для восстановления знаков точки (0.93 и 0.7) и запятой (0.92 и 0.77). При восстановлении знаков вопроса результаты приведенных моделей по F1 различались на 0.2%, составив 0.418 и 0.42, соответственно. Примечательно, что для модели [9] это обуславливалось "средними" значениями метрик Precision (0.55) и Recall (0.34), в то время как у модели [8] – высоким значением метрики Precision (0.76) при сравнительно низком значении метрики Recall (0.29).

Наличие пунктуации может само по себе помочь расстановке пауз моделью в момент синтеза, если модель обучается синтезировать паузы, исходя из скрытых представ-

лений, не используя непосредственную аннотацию [10]. Однако некоторые исследования, проведенные для китайского языка, показывают, что разметка "просодических границ", определяющихся на уровне слогов, слов и даже отдельных предложений, также может улучшить качество синтеза. Так, например, в работе "Automatic Prosody Annotation with Pre-Trained Text-Speech Model" [11] предложен иерархический подход к разметке просодических границ для китайского языка с использованием механизма кросс-внимания между аудио- и текстовой расшифровкой. Это позволяет модели выучить соответствие между промежутками звучащей речи в аудио и тексте, чтобы затем автоматически размечать просодические границы (здесь – поделенные на классы пауз на уровне отдельных иероглифов, слов и так далее). Авторы отмечают, что поскольку разметка проводилась без участия человека, система демонстрирует повышенную консистентность в сравнении с ручной разметкой, что, в свою очередь, приводит к улучшенному качеству синтезированной речи.

Для получения прямой разметки пауз по длительности могут применяться также алгоритмы детекции речевой активности (Voice Activity Detection, VAD). Модель транскрибации WhisperX [12] поддерживает аннотацию на уровне слов и может, таким образом, быть использована без модификаций даже для корпусов, изначально не содержащих транскрибацию аудиоданных. В условиях ограниченных вычислительных ресурсов на полную транскрибацию возможно также использование гибридного подхода с такими алгоритмами, как WebRTC [17], использующим вероятностный подход на основе смесей распределений Гаусса и Лапласа, или Silero VAD [18], основанный на сверточных нейронных сетях (Convolutional Neural Network, CNN). Согласно официальной документации, на восьми наборах данных зашумленной речи Silero VAD показывает медианное значение метрики ROC-AUC в 0.95. Медианное значение ROC-AUC у WebRTC на тех же данных – 0.76.

Примерный алгоритм в случае использования методов VAD будет состоять из следующих этапов:

1. Детекция речевой активности выбранным алгоритмом VAD;
2. Сопоставление временных промежутков, выделенных в шаге 1, с текстовой расшифровкой из изначального датасета;
3. Аннотация текстовой расшифровки полученными метками длительности пауз.

Второй этап может быть реализован за счет использования эвристики, например, путем проставления временных меток по тексту пропорционально длительности аудио. При таком подходе пауза, находящаяся в середине аудио, будет аннотирована как находящаяся примерно в середине текста, если текст сегментирован по словам. Данный подход является наиболее простым в реализации, однако его качество сильно зависит от равномерности темпа речи и подходит преимущественно для случаев, когда речь не содержит значительных промежутков ускорения или замедления.

Другим возможным решением может быть повторное использование модели транскрибации, но не на всем аудио, а только на временных промежутках, выделенных моделью VAD. Для сопоставления полученной транскрибации с уже имеющейся в исходном наборе данных, можно использовать временное окно, охватывающее соответствующий сегмент аудио, а также дополнительные, например, 100 миллисекунд до и после него. Сравнение текстов можно осуществлять с использованием метрик посимвольного сходства (например, сходство Левенштейна в реализации библиотеки RapidFuzz [13]). Если значение сходства для некоторого участка исходной транскрибации превышает заданный порог, то метка паузы может быть проставлена в данном отрывке текста.

2.2. Автоматическая разметка ударений

2.2.1. Подход с использованием правил (словарей)

Ударения являются одной из базовых характеристик речи и потому также представляют интерес с точки зрения автоматизации их разметки. "Наивный" подход к автоматизации разметки ударений (акцентуации) включает в себя разметку по заранее определенным правилам. Разметка в таком случае осуществляется согласно подключаемому к алгоритму словарю, содержащему информацию об ударениях в тех или иных словоформах. Данный подход является простым в реализации и не требует больших вычислительных ресурсов.

Программную реализацию данного подхода можно найти, например, у А. Полякова [14]. В документации указано, что программа способна размечать два вида ударений: первичные (в словах типа "ёлка", "база") и второстепенные (в словах типа "авиабаза"). Если одна и та же словоформа может иметь несколько ударений, то данная программа проставляет оба. Существует также возможность подключения пользовательских словарей.

Одним из минусов данного подхода является сложность определения ударений в омографах (зАмок, замОк), а также при использовании в поэтических текстах, где авторское ударение может отличаться от общепринятого. Проблема акцентуации омографов может быть частично решена с помощью расширения алгоритма и использования, например, конечных автоматов [15]. В сочетании с данными о частотности тех или иных словоформ, такой подход достиг значения метрики Accuracy в 96.15% на небольшом, вручную размеченном авторами статьи, корпусе, содержащем 7689 токенов. Данный корпус, однако, был собран из материалов, ориентированных на изучающих русский язык (диалоги, отрывки из классических произведений, а также вручную подобранные предложения). Расширение такой разметки на другие домены и масштабирование на использование большего количества данных представляется трудозатратным.

Рассмотрим подходы, позволяющие учитывать контекст и использующиеся таким образом для решения проблем с акцентуацией омографов.

2.2.2. Нейросетевой подход

Для учета контекста могут использоваться нейронные сети. В частности, такие архитектуры как рекуррентные нейронные сети (Recurrent Neural Network, RNN), их разновидности, такие как сети с долгой краткосрочной памятью (Long Short-Term Memory, LSTM), а также более продвинутые архитектуры типа "Трансформер", предназначенные для обработки последовательных данных. Представленные в 2017 году в статье "Attention is all you need" [16], сети трансформерной архитектуры обрабатывают входящие последовательности с помощью так называемого "механизма внимания". Благодаря ему стала возможной обработка не только отдельных слов и кратковременных контекстов, но также учет расширенного, по сравнению с RNN и LSTM, контекстного окна и ускорение обучения моделей за счет использования параллелизации вычислений.

Примером разметки ударений нейросетевым подходом является разработка И. Гусева [17]. В программном комплексе имеется возможность использовать разновидность модели трансформерной архитектуры deberta-2 или LSTM. Данные для обучения собираются из открытых источников (Викисловарь, "Грамматический словарь русского языка" А. А. Зализняка [18], а также вручную размеченный набор данных), а затем подаются на вход для обучения модели. В официальной документации сказано, что таким образом удалось достичь значения 89.73% по метрике Accuracy, без указания, однако, на какой из моделей данное значение было получено.

Для русскоязычных текстов сугубо нейросетевой подход исследовался в рамках разметки поэзии, не проверенной профессиональными редакторами и выложенной в открытом доступе на ресурсе stihi.ru [19]. Авторы приводят примеры некоторых удачных определений ударения, однако на момент написания статьи их код являлся недоступным для использования и модификации.

Существуют и программные средства с открытым исходным кодом, направленные на акцентуацию в области поэтических текстов. Например, программный пакет "RussianPoetryScansionTool" [20, 21] позволяет расставлять ударения в текстах, а также оценивать их стихотворный размер и рифму. Пользователю предоставляется возможность использовать на выбор одну из четырех архитектур моделей: многослойный перцептрон с ReLU-активациями, LSTM, CNN или модель трансформерной архитектуры с собственными весами. В официальной документации не приведены метрики качества работы моделей, однако есть приведенный пример использования библиотеки для аннотации стихотворения. Примечательно, что библиотека ставит ударения в словах, содержащих букву "ё", однако не всегда ставит ударение в односложных словах ("столь", "чтоб"). Также отмечено, что основные и второстепенные (при наличии) ударения обозначаются разными символами.

Наконец, благодаря использованию нейросетевой архитектуры двунаправленных LSTM, авторам статьи "Automated Word Stress Detection in Russian" [22, 23] удалось добиться микроусредненного (по классам слов, содержащих от двух до девяти слогов) значения Accuracy в 0.979 на наборе данных в 1154067 уникальных тестовых примеров. Такое значение было получено при использовании модели, учитывающей контекст в формате окончания предыдущего слова. Модель, не использовавшая данную информацию, показала себя незначительно хуже – для нее микроусредненная метрика Accuracy равнялась 0.977.

Важным фактом является то, что значение метрики Accuracy обеих моделей было значительно ниже при тестировании на пятидесяти омографах. Для модели, учитывающей контекст, она составила 0.819, а для модели, не учитывающей контекст – 0.77. Эксперименты авторов также показали, что использование данных из размеченных корпусов является предпочтительным использованию данных из словарей, поскольку в первом случае слова находятся в контексте и позволяют модели учитывать частотность возможных ударений.

Стоит отметить, что, поскольку нейросетевые подходы основываются на частотных закономерностях языка, они также подвержены изменениям в обучающей выборке и могут плохо показывать себя на примерах, слабо репрезентированных в обучающих данных, таких как авторские изменения ударения в поэтическом домене, если модель была обучена сугубо на прозаических текстах.

2.2.3. Комбинированный подход

Комбинированный подход к разметке ударений в поэтических текстах в русском языке представлен в статье "Комбинированный Словарно-Нейросетевой Акцентуатор Для Разметки Русского Поэтического Текста" [24]. Как было отмечено ранее, разметка таких текстов представляет особую сложность, поскольку наряду с омографами может содержать также и авторские ударения, обусловленные ритмикой конкретного стихотворения.

Взяв за основу разработку [22], не используя сторонние библиотеки для POS-тегинга (такие, как pymorphy или SpaCy), авторы учитывали морфологический контекст с помощью использования флексии предшествующего слова. Обучающая выборка собиралась из "Грамматического словаря русского языка" А. А. Зализняка и устного подкорпуса Национального корпуса русского языка (НКРЯ) [25]. Авторы отмечают, что

ошибки, совершаемые словарным и нейросетевым акцентуаторами, отличаются по своей сути. Словарный акцентуатор ошибается в определении ударений в словах, отсутствующих в словаре, а также в случаях с неоднозначным ударением. Двунаправленная LSTM-модель ошибалась, например, при определении ударения в словах с подразумеваемой, но не обозначенной буквой "ё", что может указывать на необходимость предварительной "ёфикации" (в наиболее простой реализации: за счет словарей) текстов перед использованием такого подхода.

Авторы рассмотрели несколько способов совмещения подходов – изначальная разметка словарным акцентуатором с последующей разметкой нейросетевым акцентуатором, изначальная разметка нейросетевым акцентуатором с последующей разметкой словарным акцентуатором, а также случайный выбор разметки тем или иным акцентуатором для каждого слова.

В итоговом пайплайне реализована одновременная разметка обоими акцентуаторами, причём результаты разметки нейросетевого акцентуатора учитываются только в случае неоднозначной разметки, отсутствия разметки словарным акцентуатором (при условии отсутствия в слове буквы "ё"), либо наличия и разметки, и буквы "ё" (кроме слов с дефисом). Результаты разработки реализованы в виде библиотеки ru-accent-poet [26] на языке Python, доступной для скачивания.

За счет использования словаря при разметке ударений для однозначных слов и использовании нейросетей для разметки ударений в омографах, удалось достичь результатов, превосходящих использование только одного из методов. Качество работы сравнивалось на вручную размеченных авторами стихах, а также на 100 строках из поэтического подкорпуса НКРЯ. На данной выборке комбинированный подход показал качество около 0.98 по метрике Accuracy, по сравнению с 0.93 у словарного и 0.94 у отдельно нейросетевого подходов, соответственно.

2.3. Автоматическая разметка высоты и тембра

Рассмотрим такие характеристики как высота и тембр голоса. Под высотой будем понимать слуховое ощущение частоты звука, а под тембром – признак, позволяющий слушателю различать звуки одинаковой высоты и громкости, но различного генезиса [27]. Высота звука тесно связана с его "фундаментальной частотой" (частота основного тона, F0), в то время как тембр – с обертонами, то есть всей частью звукового спектра, не относящейся к фундаментальной частоте [28].

Исследования показывают, что высота голоса играет ключевую роль в восприятии социальных качеств, таких как надежность, авторитетность, лидерские качества говорящего [29, 30, 31, 32]. Таким образом, моделирование и автоматическая разметка высоты и тембра являются особенно актуальными при, например, создании голосовых ассистентов, где формирование доверительного отношения пользователей является одним из наиболее важных факторов в разработке.

2.3.1. Выделение фундаментальной частоты и тональных контуров

Алгоритмы выделения фундаментальной частоты, как правило, опираются на анализ аудиосигнала во временной либо частотной области, также существуют и гибридные подходы [33]. Поскольку данная статья рассматривает разметку в целях генерации аудио, содержащего характеристики только одного говорящего, то рассматриваться будут только базовые алгоритмы, применяющиеся при определении F0 в вышеуказанном сценарии.

При работе с речью в частотной области часто применяются мел-кепстральные коэффициенты (Mel-frequency Cepstrum Coefficient, MFCC) – представление, полученное

через обратное преобразование Фурье от логарифма спектра мощности сигнала. Использование мел-шкалы позволяет учесть нелинейную связь между воспринимаемой и физической частотой, возникающую в результате особенностей человеческой физиологии.

Большинство алгоритмов, использующих преимущественно временное представление входного сигнала для определения F0, основаны на принципе автокорреляции. Входной сигнал разделяется на части (фреймы), от которых высчитывается автокорреляционная функция, отображающая сходство сигнала с самим собой. В наиболее простом подходе первый максимум данной функции и будет фундаментальной частотой. На базе этого подхода были разработаны несколько алгоритмов, таких как AMDF [34], YIN [35], а также его вероятностная модификация pYIN [36] и другие.

В алгоритме YIN используется кумулятивное нормализованное среднее, что позволяет ему быть более устойчивым к колебаниям в амплитуде входного сигнала, делая период F0 более выраженным по отношению к остальным. Программная реализация YIN и pYIN доступна в библиотеке librosa [37].

Подходы, использующие исключительно анализ в частотной области, пользуются меньшей популярностью и их реализации не настолько распространены, как реализации гибридных подходов или подходов, основанных на анализе амплитудно-временных характеристик. Среди известных гибридных подходов можно отметить HARVEST [37] и YAAPT [38].

HARVEST извлекает основные частоты-кандидаты F0 с помощью набора фильтров с разными частотами, анализируя спектральные компоненты, после чего уточняет их с использованием "мгновенной" частоты. Затем несколько кандидатов F0 оцениваются в каждом фрейме. Для формирования финального F0-контура применяется алгоритм соединения соседних кадров, опирающийся на предположение о плавности изменения высоты тона, что делает его более устойчивым к локальным шумам (проблема, проявляющаяся при покадровой обработке сигналов). Реализация алгоритма доступна в библиотеке pyworld [39], также предоставляющей реализацию алгоритма DIO [40].

Ядром алгоритма YAAPT является метод нормализованной кросс-корреляции, использующийся вместо обычной автокорреляционной функции. На этапе предобработки над входным сигналом производится нелинейное преобразование, позволяющее восстановить слабые компоненты F0. Для выбора наиболее правдоподобной F0 используется динамическое программирование, что делает алгоритм устойчивым к искажениям и эффективным при работе как с записями высокого качества, так и с, например, телефонной речью. Реализация алгоритма YAAPT доступна для Python в библиотеке AMFM_decompy [41].

Существуют также статьи, описывающие успешное применение методов традиционного машинного обучения (алгоритмов K-Среднего, модели смесей Гауссовских распределений, метода опорных векторов [42]) и CNN [43] для решения задачи выделения F0.

Помимо отдельной частоты F0 можно выделять также тональный контур, представляющий собой изменение тона на протяжении отдельного отрезка звучащей речи. Одним из наиболее распространенных программных решений для работы с анализом речи, дающим возможность выделения тональных контуров, является программный пакет Praat [44], доступный для использования с языком Python с помощью библиотеки Parselmouth [45]. Алгоритм, применяющийся в Parselmouth для выделения F0 по умолчанию, является автокорреляционным [46], однако пользователю предоставлена возможность выбрать и другие методы.

Для разметки текстовых данных, поступающих в модель на момент синтеза, можно использовать моделирование тональных контуров на основе частотных характеристик частей речи, как это было сделано для тамильского языка в статье "Utilizing POS-Driven

Pitch Contour Analysis for Enhanced Tamil Text-to-Speech Synthesis" [47], однако возможность успешного использования такого подхода применительно к русскому языку требует дополнительных исследований.

Рассмотренные выше методы в основном были направлены на разметку базовых характеристик звучащей речи – ударения, пауз и тона. Хотя их модуляция и способствует повышению естественности синтезированной речи, этого недостаточно для моделирования речевого многообразия. В следующем разделе будут рассмотрены способы автоматической разметки более сложных аспектов стиля, таких как акцент и эмоция говорящих.

2.3.2. Разметка характеристик тембра в текстовом формате

Наиболее простым с точки зрения естественного языка способом задать желаемый голос является непосредственное задание характеристик словесным образом. Ряд исследований сфокусировался на создании таких словесных описаний. Например, авторами статьи "Dream Voice: Text Guided Voice Conversion" [48] была разработана система из десяти ключевых слов, разделенных на две категории в зависимости от уровня субъективности. Первая категория составляла базовые характеристики, такие как пол и возраст говорящего, вторая же соответствовала более абстрактным понятиям, таким как сила или теплота голоса. Разметка проводилась вручную экспертами, а в итоговом датасете было 900 говорящих.

Авторы статьи [1] развили идею стилистической разметки, предложив подход к ее автоматизации. Дополнительно к этому ими была предложена обширная система тегов, охватывающих как присущие отдельным говорящим характеристики ("Intrinsic tags"), такие как акцент и пол, так и ситуативные характеристики, такие как выражаемая в речи эмоция ("Situational tags"). Стоит отметить, что для разметки ситуативных характеристик не подходит использование исключительно методов, основанных на анализе тональности текста, так как интонация высказывания может отличаться от семантики предложения.

Авторы также выделяли уровень "сложности" тегов, где "базовыми" ("Basic") считались такие теги, как пол говорящего, скорость речи и высота голоса, поддающиеся определению с помощью методов обработки сигналов, а "расширенными" ("Rich") считались такие теги, как выражаемая эмоция, акцент и другие, обычно требующие человеческой разметки.

Сфокусировавшись на масштабировании расширенных характеристик (R-тегов), авторы проанализировали существующие, преимущественно англоязычные, наборы данных (датасеты). Из находящихся в открытом доступе датасетов, 1 из 10 имел разметку исключительно присущих R-тегов [49], в 6 из 10 присутствовала разметка только ситуативных R-тегов [2, 50, 51, 52, 53, 54] и в 1 из 10 [55] – разметка как присущих, так и ситуативных R-тегов. При этом датасеты, разметка которых производилась автоматически, не имели разметки присущих R-тегов, а размеченные ситуативные R-теги ограничивались 4 [2] и 7 [54] тегами, соответственно. Таким образом, была выявлена необходимость автоматизации разметки R-тегов.

Дальнейшая разметка производилась двумя способами, отдельно для присущих R-тегов (IR-тегов) и ситуативных R-тегов (SR-тегов). Пайплайн для IR-тегов начинался с ручной разметки небольшого "стартового" датасета. Затем датасет масштабировался путем нахождения голосов, похожих на известные, и переноса на них имеющихся IR-тегов. Для каждого размеченного вручную голоса и каждого голоса из размечаемого датасета, авторы вычисляли медианные эмбеддинги на основе десяти случайно выбранных аудиофрагментов, используя модель VoxSim [56]. Особенностью данной модели является то, что она обучалась определять не то, насколько разные фрагменты голосов при-

надлежат одному и тому же голосу, но то, насколько те или иные фрагменты воспринимаются похожими с точки зрения человека ("perceptual speaker similarity"). Авторы статьи отметили, что если два голоса имеют высокое сходство по восприятию, то у них обычно совпадает большинство IR-тегов.

Для каждого говорящего из размеченного датасета находились говорящие из размечаемого датасета, косинусное сходство с которыми было больше или равно 0.8, а затем размечаемому говорящему копировались все IR-теги.

Вторая часть пайплайна, использовавшаяся для разметки SR-тегов, состояла из трех этапов. На первом из них авторы фильтровали наиболее эмоционально окрашенные высказывания, пользуясь готовым классификатором для трехфакторного моделирования эмоциональных состояний по шкалам степени контроля (dominance), интенсивности (arousal) и приятности (valence) [57]. Отметим, что в официальной документации модели описано ее применение для классификации эмоций в аудио-текстовом наборе данных "The Berlin Database of Emotional Speech" [58] на немецком языке, содержащем разметку аудио на пять эмоций, плюс одну "нейтральную" эмоцию. Обученный на основе эмбеддингов данной модели классификатор, использующий метод опорных векторов (Support Vector Classifier, SVC), показал значение метрики Unweighted Average Recall (UAR) в 0.93, что говорит в пользу ее применимости для языков, отличных от английского.

На втором этапе текстовая расшифровка оценивалась на предмет соответствия семантики размечаемой характеристике. С помощью модели SFR-Embedding-Mistral [59], авторы вычисляли косинусное сходство между промптом: "Instruct: Given an emotion, retrieve relevant transcript lines whose overall style/emotions matches the provided emotion. Query: {emotion}" ["Инструкция: По заданной эмоции верни релевантные строки расшифровки, стиль/эмоция которых соответствует предложенной. Запрос: {Эмоция}"] и текстовыми расшифровками речи, полученными в результате первого этапа фильтрации. Чтобы избежать переоценки реплик, в которых просто упоминается эмоция (например, реплики, содержащие слово "ярость", но не имеющие соответствующего эмоционального окраса), реплики также фильтровались по ключевым словам. После получения косинусного сходства реплики ранжировались от наиболее до наименее подходящих под промпт.

Заключительным этапом было акустическое сопоставление. Для фильтрации ложноположительных срабатываний после второго этапа, авторы брали топ-100 тысяч реплик, наиболее подходивших под промпт той или иной эмоции. Выбранные реплики подавались на вход аудио-модели Gemini 1.5 Flash. Модель оценивала, насколько интонация соответствует заданной эмоции по шкале от 1 до 5, при этом в промпте содержалось указание не учитывать семантику высказывания. В результате оставлялись только реплики, получившие оценку 5. Результатом данной части пайплайна стал набор реплик, точно отражающих нужную эмоцию и по содержанию, и по звучанию. Проведенное исследование с удалением части компонент ("ablation study") пайплана показало, что каждый из них необходим для получения более качественного результата.

В целях оценки возможности применимости такой модели фильтрации к русскоязычным аудио, авторами данной статьи также был проведен мини-эксперимент по ее использованию. Для сравнения было выбрано высказывание "Я так люблю эту жизнь, я самый счастливый человек на планете Земля", произнесенное с грустной эмоцией, прямо противоположной его ярко-выраженной положительной семантике. Нескольким моделям серии Gemini было предложено поставить оценку от 1 до 5, сравнив, насколько выражаемая в аудио эмоция соответствует эмоции "грусть". Использовавшийся промпт соответствовал промпту из статьи [1], однако название эмоции указывалось на русском: "Analyze the provided speech clip to evaluate how effectively it conveys the emotion {emotion

→ Грусть}, focusing on tone of voice and delivery, rather than the spoken content... " ["Проанализируй данный отрывок речи для того, чтобы оценить, насколько эффективно он передает эмоцию {эмоция → Грусть}, фокусируясь на тоне голоса и подаче, нежели на содержании высказывания...}"].

Модель "gemini-1.5-flash-002", применявшаяся в изначальной статье, не справилась с поставленной задачей, поставив оценку 1/5. Однако более новая модель "gemini-2.5-flash-preview-04-17" в режиме "Thinking mode" справилась с задачей, поставив оценку 5/5, что указывает на потенциал использования ее API для реализации аналогичного пайплайна для русскоязычных данных.

3. Результаты

В статье были проанализированы основные подходы к автоматизации разметки паралингвистических характеристик речи, таких как паузы, ударения, высота, а также "R-теги", включающие эмоции, акцент и другие особенности говорящих.

Наиболее исследованным из направлений автоматизации разметки является разметка ударений, в частности для домена поэтических текстов. Устоявшиеся алгоритмы к моделированию просодических границ и тональных контуров могут быть использованы без дополнительных модификаций. Существуют также нестандартные подходы, успешно применявшиеся для китайского и тамильского языков, однако требующие апробации для доказательства эффективности на русскоязычных данных. В силу различной природы происхождения текстовых данных, для повышения качества и обеспечения стабильности работы алгоритмов, рекомендуется использовать предобработку, включающую в себя "ёфикацию", капитализацию и восстановление знаков препинания.

Сводная информация о доступных методах автоматической разметки базовых характеристик на русскоязычных данных представлена в таблице.

Сводная таблица доступных программных решений для разметки базовых паралингвистических характеристик на русском языке

Размечаемая характеристика	Программное решение
Паузы	Для восстановления знаков препинания: ru_punct [8] ru-autopunctuation [9]
	Для транскрибации и временных меток: WhisperX [12]
	Для гибридного использования с моделями транскрибации: WebRTC [60] Silero VAD [61]
Ударения	Accenter [14] russ [17] RussianPoetryScansionTool [20] russtress [23] ru-accent-poet [26]
F0, Тональный контур	librosa [62] pYAAPT [41] Parselmouth [45]

В целом можно сказать, что задача автоматизации разметки базовых паралингвистических характеристик в русском языке может быть успешно решена с использованием комбинации существующих программных решений. Выбор библиотек для реализации будет зависеть от существующих вычислительных мощностей, объема обрабатываемых данных, а также требуемой для конкретной задачи разметки точности.

Так, для быстрого прототипирования подойдут методы, основанные на правилах и эвристиках. В частности – разметка пауз на основе синтаксиса или словарный подход в случае разметки ударений. Для задач, требующих повышенной точности, подойдет использование нейросетевых и гибридных методов, таких как WhisperX для разметки пауз и ru-accent-poet для разметки ударений. Стоит отметить, что все описанные методы разметки ударений требуют наличия текстовой расшифровки аудио. В условиях отсутствия такой расшифровки, с учетом особенностей ударения в русском языке, разметка может быть произведена за счет выделения участков локального изменения длительности и тембральных характеристик гласных [63] непосредственно из аудио. На момент написания статьи авторам не известно о существовании готовых программных реализаций подобных алгоритмов, что предполагает дополнительные временные затраты на их разработку.

В случае разметки расширенных характеристик наиболее популярными являются подходы, основанные на нейросетевых моделях, опирающиеся, как правило, на решения, обученные преимущественно на англоязычных данных. Например, модель VoxSim, описанная в статье [36], была обучена исключительно на англоязычном материале и может оказаться непригодной для сравнения голосов в русскоязычной речи. С другой стороны, для моделей, не обучавшихся в явном виде использовать семантику высказывания при определении эмоций, существует потенциал применимости для языков, не входивших в обучающую выборку.

В совокупности результаты анализа показывают необходимость разработки программных решений, позволяющих автоматизировать разметку расширенных характеристик для русскоязычных данных.

4. Обсуждение

В перспективе разработки методов автоматизации разметки расширенных тегов для русскоязычных данных, актуальной задачей остается и разработка собственной системы тегов. Существующие аудио-текстовые наборы данных на русском языке, такие как Dusha, resd_annotated, CommonVoice 21.0, ПРуД, РИНКО [4, 5, 6, 64, 65] и другие, не имеют единой стандартизированной системы аннотации. Они также различаются по качеству записей и квалификации разметчиков (от собранных и размеченных пользователями сети Интернет до записанных и размеченных профессионально). Некоторые из них содержат отдельные базовые или расширенные характеристики: например, в CommonVoice аннотирован пол говорящих, в ПРуД – их диалекты. При этом не все из описанных корпусов находятся в открытом доступе. Для создания системы, охватывающей как можно более гибкий спектр характеристик, необходимо провести дополнительное исследование, которое позволит определить, какие данные доступны и пригодны для использования.

Исходя из существующих решений для разметки расширенных тегов можно отметить, что некоторые из них показывают возможность масштабирования на языки, не входившие в обучающую выборку. Однако для однозначных выводов требуется проведение тестов на наборах вариативных русскоязычных данных. При интерпретации результатов стоит учитывать такие факторы, как культурные отличия в выражении тех или иных эмоций, спонтанность размечаемой речи, а также, для анализа качества работы моделей на подготовленной речи – наличие навыков сценического мастерства у говорящих.

Дополнительным направлением для улучшения системы генерации естественной речи может стать разметка и синтез экстралингвистических компонент речи – таких как смех, кашель, цоканье и другие [66]. Кроме того, перспективным является задание стиля не только говорящего, но и окружающей среды, в том числе с помощью изображений [67, 68], а также расширение на ситуации с несколькими говорящими или переключением кодов.

Данная работа ограничена по охвату и не претендует на исчерпывающее рассмотрение всех существующих подходов и методов. Авторы стремились сделать акцент на практическом применении программных решений, что, как они надеются, может способствовать прикладному развитию в области создания и масштабирования аудио-текстовых наборов русскоязычных данных.

5. Заключение

В данной статье приведен обзор некоторых из существующих методов автоматической разметки паралингвистических характеристик, а также оценен потенциал их использования для русскоязычных аудио-текстовых наборов данных. Анализ выявил, что существующих программных решений достаточно для выделения базовых характеристик, таких как паузы, ударения, фундаментальная частота и тональные контуры. Описанные методы могут использоваться для унификации и масштабирования разметки наборов данных, находящихся в открытом доступе, в целях обучения моделей синтеза речи.

Вместе с тем обнаружена нехватка методов, позволяющих автоматизировать разметку расширенных характеристик для русскоязычных данных. По этой причине возникает необходимость в адаптации и проверке методов, продемонстрировавших эффективность на англоязычных данных, с использованием русскоязычных данных и моделей.

Для дальнейшего развития в этом направлении предлагается:

1. Разработать собственную систему тегов, учитывающую как базовые, так и расширенные характеристики;
2. Провести комплексный анализ существующих русскоязычных аудио-текстовых наборов данных с точки зрения доступности и характера аннотации;
3. Выполнить сравнительный анализ методов оценки голосовой схожести, исследовать потенциал использования моделей, обучавшихся на англоязычных данных, для сравнения русскоязычной речи.

Список источников

1. *Diwan A., Zheng Z., Harwath D., Choi E.* Rich Style-Prompted Text-to-Speech Datasets, 2025. URL: http://arxiv.org/abs/2503.04713 (дата обращения: 31.03.2025).
2. *Guo Z., Leng Y., Wu Y., et al.* PromptTTS: Controllable Text-to-Speech with Text Descriptions // Proceedings of ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023. P. 1–5. DOI: 10.1109/ICASSP49357.2023.10096285.
3. *Lacombe Y., Srivastav V., Gandhi S.* Inference and training library for high-quality TTS models. URL: https://github.com/huggingface/parler-tts (дата обращения: 15.04.2025).
4. *Kondratenko V., Sokolov A., Karpov N. et al.* Large Raw Emotional Dataset with Aggregation Mechanism, 2022. URL: http://arxiv.org/abs/2212.12266 (дата обращения: 31.03.2025).
5. *Князев С. В., Мороз Г. А., Дьяченко С. В.* Корпус Просодии Русских Диалектов (ПРуД). URL: https://lingconlab.github.io/PRuD/ (дата обращения: 10.04.2025).

6. *Кривнова О. Ф., Архипов А. В., Захаров Л. М., Кобозева И. М.* Интонация устного дискурса: русский интонационный корпус РИНКО (RINCO) // Речевые Технологии. 2020. № 1–2. С. 113–120.
7. *Речевые* хезитации: формальный и функциональный аспекты / Яковлева Э. Б.: Институт научной информации по общественным наукам РАН, 2016. 74 с. URL: https://elibrary.ru/item.asp?id=30706219 (дата обращения: 21.04.2025).
8. *Гутник Г.* Нейронная сеть для восстановления пунктуации на русском языке. URL: https://github.com/gleb-skobinsky/ru_punct (дата обращения: 20.04.2025).
9. *Котик К.* Punctuation and casing restoration for the Russian Language (BERT-based). URL: https://github.com/kotikkonstantin/ru-autopunctuation (дата обращения: 20.04.2025).
10. *Hwang J.-S., Lee S.-H., Lee S.-W.* PauseSpeech: Natural Speech Synthesis via Pre-trained Language Model and Pause-Based Prosody Modeling // Pattern Recognition / eds. H. Lu et al. Springer Nature, 2023. P. 415–427.
11. *Dai Z., Yu J., Wang Y. et al.* Automatic Prosody Annotation with Pre-Trained Text-Speech Model // Proceedings of Interspeech 2022. P. 5513–5517. DOI: 10.21437/Interspeech.2022-10005.
12. *Bain M., Huh J., Han T., Zisserman A.* WhisperX: Time-Accurate Speech Transcription of Long-Form Audio // Proceedings of Interspeech. 2023. P. 4489–4493. DOI: 10.21437/Interspeech.2023-78.
13. *RapidFuzz* 3.13.0 documentation. URL: https://rapidfuzz.github.io/RapidFuzz/index.html (дата обращения: 20.04.2025).
14. *Поляков А.* Accenter/Accenter.txt at master. Accenter created by Alexey Polyakov, GitHub. URL: https://github.com/sStress/Accenter/blob/master/Accenter.txt (дата обращения: 11.04.2025).
15. *Reynolds R., Tyers F.* Automatic word stress annotation of Russian unrestricted text // Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015). 2015. P. 173–180. URL: https://aclanthology.org/W15-1822/ (дата обращения: 11.04.2025).
16. *Vaswani A. et al.* Attention is All you Need // Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017. Vol. 30. URL: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (дата обращения: 11.04.2025).
17. *Гусев И.* Package for word stress detection URL: https://github.com/IlyaGusev/russ (дата обращения: 11.04.2025).
18. *Грамматический* словарь русского языка / Зализняк А. А. М., 1977. URL: https://gramdict.ru/ (дата обращения: 11.04.2025).
19. *Гришина Е. А., Зеленков Ю. Г., Орехов Б. В.* Наивная Поэзия В Акцентологическом Корпусе // Труды Института русского языка им. В. В. Виноградова. 2015. Т. 3, № 6. С. 257–272.
20. *Koziev I.* Detection of poetic meter, rhyme, and stress placement in the texts of Russian accentual-syllabic poems and songs URL: https://github.com/RussianNLP/RussianPoetryScansionTool (дата обращения: 13.04.2025).
21. *Koziev I.* Automated Evaluation of Meter and Rhyme in Russian Generative and Human-Authored Poetry, 2025. URL: http://arxiv.org/abs/2502.20931 (дата обращения: 16.04.2025).
22. *Ponomareva M., Milintsevich K., Chernyak E., Starostin A.* Automated Word Stress Detection in Russian // Proceedings of the First Workshop on Subword and Character Level Models in NLP SCLeM. 2017. P. 31–35. URL: https://aclanthology.org/W17-4104/ (дата обращения: 11.04.2025).

23. *Ponomareva M.* Python package russtress accentuates russian text. URL: https://github.com/MashaPo/russtress (дата обращения: 13.04.2025).
24. *Короткова Ю. О.* Комбинированный Словарно-Нейросетевой Акцентуатор Для Разметки Русского Поэтического Текста // Труды Института русского языка им. В. В. Виноградова. 2022. № 3. С. 181–190.
25. *Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А. и др.* Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы Языкознания. 2024. № 2. С. 7–34.
26. *Короткова Ю. О.,* ru-accent-poet. URL: https://github.com/yuliya1324/ru_accent (дата обращения: 11.04.2025).
27. *Педагогическое* речеведение: словарь-справочник под ред. Т. А. Ладыженской и А. К. Михальской / Князьков А. А. под ред. Т. А. Ладыженской и А. К. Михальской. 1998. URL: http://rus-yaz.niv.ru/doc/pedagogical-speech/index.htm (дата обращения: 11.04.2025).
28. *Pitch* Extraction and Fundamental Frequency: History and Current Techniques. Pitch Extraction and Fundamental Frequency / Gerhard D. Department of Computer Science, University of Regina, 2003. 44 p.
29. *Klofstad C. A. et al.* Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women // Proceedings in Biological Sciences. 2012. Vol. 279, № 1738. P. 2698–2704.
30. *O'Connor J. J. M. et al.* The influence of voice pitch on perceptions of trustworthiness across social contexts // Evolution and Human Behavior. 2017. Vol. 38, № 4. P. 506–512.
31. *Wang T.-Y., Kawaguchi I., Kuzuoka H., Otsuki M.* Effect of Manipulated Amplitude and Frequency of Human Voice on Dominance and Persuasiveness in Audio Conferences // Proc. ACM Human-Computer Interaction. 2018. Vol. 2. № CSCW. P. 177:1–177:18.
32. *Wu H. X., Li Y., Ching B. H-H., Chen. T. T.* You are how you speak: The roles of vocal pitch and semantic cues in shaping social perceptions // Perception. 2023. Vol. 52, № 1. P. 40–55.
33. *Имамвердиев Я. Н., Сухостат Л. В.,* Подходы для оценки периода основного тона речевого сигнала в зашумлённой среде // Речевые Технологии. 2014. № 1–2. С. 84–103.
34. *Ross M., Shaffer H., Cohen A. et al*. Average magnitude difference function pitch extractor // IEEE Transactions on Acoustics, Speech, and Signal Processing. 1974. Vol. 22, № 5. P. 353–362.
35. *De Cheveigné A., Kawahara H.* YIN, a fundamental frequency estimator for speech and music // The Journal of the Acoustical Society of America. 2002. Vol. 111, № 4. P. 1917–1930.
36. *Mauch M., Dixon S.* PYIN: A fundamental frequency estimator using probabilistic threshold distributions // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. P. 659–663. URL: https://ieeexplore.ieee.org/document/6853678 (дата обращения: 15.04.2025).
37. *Morise M.* Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals // Proceedings of Interspeech 2017. P. 2321–2325. DOI:10.21437/Interspeech.2017-68.
38. *Kasi K., Zahorian S. A.* Yet Another Algorithm for Pitch Tracking // 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. 2002. Vol. 1. P. I-361-I-364. URL: https://ieeexplore.ieee.org/document/5743729 (дата обращения: 15.04.2025).

39. *Hsu J. et al.* PyWorld: a Python wrapper for WORLD vocoder. URL: https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder (дата обращения: 20.04.2025).
40. *Morise M., Kawahara H., Katayose H.* Fast and Reliable F0 Estimation Method Based on the Period Extraction of Vocal Fold Vibration of Singing Voice and Speech // CD-ROM Proceeding AES 35th International Conference: Audio for Games. London, United Kingdom, 2009.
41. *Schmitt B. J. B.* pYAAPT. AMFM_decompy 1.0.11 documentation. URL: https://bjbschmitt.github.io/AMFM_decompy/pYAAPT.html (дата обращения: 15.04.2025).
42. *Drugman T., Huybrechts G., Klimkov V., Moinet A*. Traditional Machine Learning for Pitch Detection // IEEE Signal Processing Letters. 2018. Vol. 25, № 11. P. 1745–1749.
43. *Kim J. W., Salamon J., Li P., Bello J. P.* CREPE: A Convolutional Representation for Pitch Estimation. CREPE // Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. P. 161–165. DOI: 10.1109/ICASSP.2018.8461329.
44. *Boersma P., van Heuven P.* Praat, a system for doing phonetics by computer // Glot International. 2001. Vol. 5, № 9/10. P. 341–345.
45. *Jadoul Y., Thompson B., de Boer B.* Introducing Parselmouth: A Python interface to Praat // Journal of Phonetics. 2018. Vol. 71. P. 1–15.
46. *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. 1993. Vol. 17, № 1193. P. 97–110.
47. *Thinakaran P., Gladston A. R., Vijayalakshmi P. et al.* Utilizing POS-Driven Pitch Contour Analysis for Enhanced Tamil Text-to-Speech Synthesis // Proceedings of the 21st International Conference on Natural Language Processing (ICON). 2024. P. 269–273.
48. *Hai J., Thakkar K., Wang H. et al.* DreamVoice: Text-Guided Voice Conversion // Proceedings of Interspeech 2024. P. 4373–4377. DOI: 10.21437/Interspeech.2024-1432.
49. *Kawamura M., Yamamoto R., Shirahata Y. et al.* LibriTTS-P: A Corpus with Speaking Style and Speaker Identity Prompts for Text-to-Speech and Style Captioning. // Proceedings of Interspeech 2024. P. 1850–1854. DOI: 10.21437/Interspeech.2024-692.
50. *Nguyen T. A., Hsu W.-N., D'Avirro A. et al.* EXPRESSO: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis // Proceedings of Interspeech 2023. P. 4823–4827. DOI: 10.21437/Interspeech.2023-1905.
51. *Richter J., Wu Y.-C., Krenn S., et al.* EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation // Proceedings of Interspeech 2024. P. 4873–4877. DOI: 10.21437/Interspeech.2024-153.
52. *Ji S., Zuo J., Fang M. et al.* TextrolSpeech: A Text Style Control Speech Corpus with Codec Language Text-to-Speech Models // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. P. 10301–10305. URL: https://ieeexplore.ieee.org/document/10445879 (дата обращения: 31.03.2025).
53. *Guan W., Li Y., Li T.et al.* MM-TTS: Multi-modal Prompt based Style Transfer for Expressive Text-to-Speech Synthesis // Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. P. 18117–18125. DOI: 10.1609/aaai.v38i16.29769.

54. *Jin Z. et al.* SpeechCraft: A Fine-Grained Expressive Speech Dataset with Natural Language Description // Proceedings of the 32nd ACM International Conference on Multimedia. 2024. P. 1255–1264. DOI: 10.1145/3664647.3681674.
55. *Watanabe A., Takamichi S., Saito Y., et al.* Coco-Nut: Corpus of Japanese Utterance and Voice Characteristics Description for Prompt-based Control // 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). P. 1–8. DOI: 10.1109/ASRU57964.2023.10389693.
56. *Nagrani A., Chung J. S., Xie W., Zisserman A.* Voxceleb: Large-scale speaker verification in the wild // Computer Speech & Language. 2020. Vol. 60. P. 101027. DOI: 10.1016/j.csl.2019.101027.
57. *Wagner J., Triantafyllopoulos A., Wierstorf H. et al.* Dawn of the transformer era in speech emotion recognition: closing the valence gap // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023. Vol. 45, № 9. P. 10745–10759. DOI: 10.1109/TPAMI.2023.3263585.
58. *Sendlmeier W., Burkhardt F., Kienast M., Paeschke A., Weiss B.* The Berlin Database of Emotional Speech. URL: http://emodb.bilderbar.info/docu/ (дата обращения: 09.06.2025).
59. *Meng R. et al.* SFR-Embedding-Mistral:Enhance Text Retrieval with Transfer Learning. URL: https://huggingface.co/Salesforce/SFR-Embedding-Mistral (дата обращения: 15.04.2025).
60. *Wiseman J.* Python interface to the WebRTC Voice Activity Detector. URL: https://github.com/wiseman/py-webrtcvad (дата обращения: 15.04.2025).
61. *Silero* VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. URL: https://github.com/snakers4/silero-vad (дата обращения: 15.04.2025).
62. *McFee B. et al.* librosa/librosa: 0.11.0. 2025. DOI: 10.5281/zenodo.15006942.
63. *Кедрова Г. Е., Потапов В. В., Егоров А. М., Омельянова Е. Б.* Акцентология. Ударение и фонетическое оформление слова. Фонетическая природа ударных гласных // Русская фонетика: учеб. материалы. 2002. URL: https://www.philol.msu.ru/~fonetica/akcent/phon_priroda/index.html (дата обращения: 09.06.2025).
64. *Lubenets I., Davidchuk N., Amenets A.* Emotions recognition from audio and text files URL: https://huggingface.co/datasets/Aniemore/resd_annotated (дата обращения: 15.04.2025).
65. *Ardila R., Branson M., Davis K., et al.* Common Voice: A Massively-Multilingual Speech Corpus. Common Voice. // Proceedings of the Twelfth Language Resources and Evaluation Conference. P. 4218–4222. ISBN: 979-10-95546-34-4.
66. *Поволоцкая А. А., Карпов А. А.* Аналитический обзор методов автоматического анализа экстралингвистических компонентов спонтанной речи // Информатика и автоматизация. 2024. Т. 23, № 1. С. 5–38.
67. *Lee Y., Yeon I., Nam J., Chung J. S.*VoiceLDM: Text-to-Speech with Environmental Context. VoiceLDM. 2023. URL: http://arxiv.org/abs/2309.13664 (дата обращения: 15.04.2025).
68. *Jung J., Ahn J., Jung C.* VoiceDiT: Dual-Condition Diffusion Transformer for Environment-Aware Speech Synthesis. 2024. URL: http://arxiv.org/abs/2412.19259 (дата обращения: 13.04.2025).

References

1. Diwan, A., Zheng, Z., Harwath, D. and Choi, E. (2025), "Scaling Rich Style-Prompted Text-to-Speech Datasets", DOI: 10.48550/arXiv.2503.04713.

2. Guo, Z., Leng, Y., Wu, Y., Zhao, S. and Tan, X. (2022), "PromptTTS: Controllable Text-to-Speech with Text Descriptions", *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece*, pp. 1–5, DOI: 10.1109/ICASSP49357.2023.10096285.
3. Lacombe, Y., Srivastav, V. and Gandhi, S., Inference and training library for high-quality TTS models, available at: https://github.com/huggingface/parler-tts (Accessed 4.15.25).
4. Kondratenko, V., Sokolov, A., Karpov, N., Kutuzov, O., Savushkin, N. and Minkin, F. (2022), "Large Raw Emotional Dataset with Aggregation Mechanism", DOI: 10.48550/arXiv.2212.12266.
5. Knyazev, S. V., Moroz, G. A. and Dyachenko, S. V. "Russian Speech Prosody Corpus (PRuD)", available at: https://lingconlab.github.io/PRuD/ (Accessed 4.10.25).
6. Krivnova, O. F., Arkhipov, A. V., Zakharov, L. M. and Kobozeva, I. M. (2020), "Russian Discourse Intonation: RINCO – a Russian Speech Intonation Corpus", *Rechevye Tekhnologii*, no. 1-2, pp. 113–120, DOI: 10.58633/2305-8129_2020_1-2_113.
7. Yakovleva, E. B. (2016), *Rechevye khezitacii: formal'nyj i funkcional'nyj aspekty.* [Speech hesitations: formal and functional aspects], RAS Institute of Scientific Information for Social Sciences (INION), Moscow, Russia.
8. Gutnik, G. (2024), "Neural network for punctuation restoration in Russian Language", available at: https://github.com/gleb-skobinsky/ru_punct (Accessed 4.20.25).
9. Kotik, K. (2025), "Punctuation and casing restoration for the Russian Language (BERT-based)", available at: https://github.com/kotikkonstantin/ru-autopunctuation (Accessed 4.20.25).
10. Hwang, J.-S., Lee, S.-H. and Lee, S.-W. (2023), "PauseSpeech: Natural Speech Synthesis via Pre-trained Language Model and Pause-Based Prosody Modeling", *Pattern Recognition. Springer Nature*, pp. 415–427. DOI: 10.1007/978-3-031-47634-1_31.
11. Dai, Z. *et al.* (2022), "Automatic Prosody Annotation with Pre-Trained Text-Speech Model", *Proceedings of Interspeech 2022*, Incheon, Korea, pp. 5513–5517, DOI: 10.21437/Interspeech.2022-10005.
12. Bain, M., Huh, J., Han, T. and Zisserman, A. (2023), "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio", *Proc. of Interspeech 2023*, Dublin, Ireland, pp. 4489–4493, DOI: 10.21437/Interspeech.2023-78.
13. RapidFuzz 3.13.0 documentation, available at: https://rapidfuzz.github.io/RapidFuzz/index.html (Accessed 4.20.25).
14. Polyakov, A., *Accenter/Accenter.txt at master. Accenter created by Alexey Polyakov*, GitHub, available at: https://github.com/sStress/Accenter/blob/master/Accenter.txt (Accessed 4.11.25).
15. Reynolds, R.and Tyers, F. (2015), "Automatic word stress annotation of Russian unrestricted text", *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pp. 173–180.
16. Vaswani, A. *et al.* (2017). "Attention is All you Need", *Advances in Neural Information Processing Systems 2017.*
17. Gusev, I. (2019), Package for word stress detection, available at: https://github.com/IlyaGusev/russ (Accessed 4.11.25).
18. Zaliznyak, A. A. (1977), *Grammaticheskij slovar' russkogo yazyka* [Grammatical dictionary of Russian language], Moscow, available at: https://gramdict.ru/ (Accessed 11.04.2025).
19. Grishina, E. A., Zelenkov, Y. G. and Orekhov, B. V. (2015), "Naive Poetry in an Accentological Corpus", *Proceedings of the V. V. Vinogradov Russian Language Institute*, Vol. 3, no. 6, pp. 257–272.

20. Koziev, I. (2025), Detection of poetic meter, rhyme, and stress placement in the texts of Russian accentual-syllabic poems and songs, available at: https://github.com/RussianNLP/RussianPoetryScansionTool (accessed 4.13.25).
21. Koziev, I. (2025), "Automated Evaluation of Meter and Rhyme in Russian Generative and Human-Authored Poetry", DOI: 10.48550/arXiv.2502.20931.
22. Ponomareva, M., Milintsevich, K., Chernyak, E. and Starostin, A. (2017), "Automated Word Stress Detection in Russian", *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pp. 31–35, DOI: 10.18653/v1/W17-4104.
23. Ponomareva, M. (2025), Python package russtress accentuates russian text, available at: https://github.com/MashaPo/russtress (Accessed 4.13.25).
24. Korotkova, Y. O. (2022), "Combined Dictionary-Neural Network Accentuator for Annotation of Russian Poetic Text", *Proceedings of the V. V. Vinogradov Russian Language Institute*, Vol. 3, no. 33, pp. 181–190, DOI: 10.31912/pvrli-2022.3.11.
25. Savchuk, S. O., Arkhangel'skij, T. A., Bonch-Osmolovskaya, A. A., Donina, O. V., Kuznecova, Y. N., Lyashevskaya, O. N., Orekhov, B. V. and Podryadchikova, M. V. (2024), "National Russian Language Corpus 2.0: new possibilities and developmental prospects", *Topics in the study of language*, pp. 7–34, DOI: 10.31857/0373-658X.2024.2.7-34.
26. Korotkova, Y. O. (2022), ru-accent-poet, a tool for putting stress marks in russian poetic texts, available at: https://github.com/yuliya1324/ru_accent (Accessed 4.11.25).
27. Knyaz'kov, A. A. (1998), *Pedagogicheskoe rechevedenie. Slovar'-spravochnik* [Pedagogical speech studies. Reference dictionary] in T.A. Ladyzhenskoj, and A.K. Mikhal'skoj (ed), available at: http://rus-yaz.niv.ru/doc/pedagogical-speech/index.htm (Accessed 11.04.2025).
28. Gerhard, D. (2003), *Pitch Extraction and Fundamental Frequency: History and Current Techniques*, Department of Computer Science, University of Regina.
29. Klofstad, C. A., Anderson, R. C. and Peters, S. (2012), "Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women", *Proceedings in Biological Sciences*, Vol. 279, № 1738, pp. 2698–2704, DOI: 10.1098/rspb.2012.0311.
30. O'Connor, J. J. M. and Barclay, P. (2017), "The influence of voice pitch on perceptions of trustworthiness across social contexts", *Evolution and Human Behavior, 2017*, Vol. 38, no. 4, pp. 506–512, DOI: 10.1016/j.evolhumbehav.2017.03.001.
31. Wang, T.-Y., Kawaguchi, I., Kuzuoka, H. and Otsuki, M. (2018), "Effect of Manipulated Amplitude and Frequency of Human Voice on Dominance and Persuasiveness in Audio Conferences", *Proc. ACM Human-Computer Interaction*, Vol. 2, pp. 177:1–177:18, DOI: 10.1145/3274446.
32. Wu, H. X., Li, Y., Ching, B. H.-H. and Chen, T. T. (2023), "You are how you speak: The roles of vocal pitch and semantic cues in shaping social perceptions", *Perception*, Vol. 52, no. 1, pp. 40–55, DOI: 10.1177/03010066221135472.
33. Imamverdiev, Y. N. and Sukhostat, L. V. (2014), "Approaches for estimating fundamental pitch period of speech signal in a noisy environment", *Rechevye Tekhnologii*, no. 1-2, pp. 84–103.
34. Ross, M. *et al.* (1974), "Average magnitude difference function pitch extractor", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 22, no. 5, pp. 353–362, DOI: 10.1109/TASSP.1974.1162598.
35. De Cheveigné, A. and Kawahara, H. (2002), "YIN, a fundamental frequency estimator for speech and music", *The Journal of the Acoustical Society of America*, Vol. 111, no. 4, pp. 1917–1930, DOI: 10.1121/1.1458024.

36. Mauch, M. and Dixon, S. (2014), "PYIN: A fundamental frequency estimator using probabilistic threshold distributions", *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 659–663, DOI: 10.1109/ICASSP.2014.6853678.
37. Morise, M. (2017), "Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals", *Proceedings of Interspeech 2017*, pp. 2321–2325, DOI: 10.21437/Interspeech.2017-68.
38. Kasi, K. and Zahorian, S. A. (2002), "Yet Another Algorithm for Pitch Tracking", *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I-361–I-364, DOI: 10.1109/ICASSP.2002.5743729.
39. Hsu J. et al., PyWorld: a Python wrapper for WORLD vocoder, available at: https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder (Accessed 4.20.25).
40. Morise, M., Kawahara, H. and Katayose, H. (2009), "Fast and Reliable F0 Estimation Method Based on the Period Extraction of Vocal Fold Vibration of Singing Voice and Speech", *Proceedings of the AES 35th International Conference: Audio for Games,* London, United Kingdom.
41. Schmitt, B. J. B., pYAAPT — AMFM_decompy 1.0.11 documentation, available at: https://bjbschmitt.github.io/AMFM_decompy/pYAAPT.html (Accessed 4.15.25).
42. Drugman, T., Huybrechts, G., Klimkov, V. and Moinet, A. (2018), "Traditional Machine Learning for Pitch Detection", *IEEE Signal Processing Letters*, Vol. 25, no. 11, pp. 1745–1749, DOI: 10.1109/LSP.2018.2874155
43. Kim, J. W., Salamon, J., Li, P. and Bello, J. P. (2018), "CREPE: A Convolutional Representation for Pitch Estimation", *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165, DOI: 10.1109/ICASSP.2018.8461329.
44. Boersma, P., and van Heuven, V. (2001), "Praat, a system for doing phonetics by computer", *Glot International*, Vol. 5, no. 9/10, pp. 341–345.
45. Jadoul, Y., Thompson, B. and de Boer, B. (2018), "Introducing Parselmouth: A Python interface to Praat", *Journal of Phonetics*, Vol. 71, pp. 1–15, DOI: 10.1016/j.wocn.2018.07.001.
46. Boersma, P. (1993), "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*, Vol. 17, no. 1193, pp. 97–110.
47. Thinakaran, P. *et al.* (2024), "Utilizing POS-Driven Pitch Contour Analysis for Enhanced Tamil Text-to-Speech Synthesis", *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, 2024, pp. 269–273.
48. Hai, J., Thakkar, K., Wang, H., Qin, Z. and Elhilali, M. (2024), "DreamVoice: Text-Guided Voice Conversion", *Proceedings of Interspeech 2024*, pp. 4373–4377, DOI: 10.21437/Interspeech.2024-1432.
49. Kawamura, M., Yamamoto, R., Shirahata, Y., Hasumi, T. and Tachibana, K. (2024), "LibriTTS-P: A Corpus with Speaking Style and Speaker Identity Prompts for Text-to-Speech and Style Captioning", *Proceedings of Interspeech 2024*, pp. 1850–1854, DOI: 10.21437/Interspeech.2024-692.
50. Nguyen, T. A. *et al.* (2023), "EXPRESSO: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis", *Proceedings of Interspeech 2023*, pp. 4823–4827. DOI: 10.21437/Interspeech.2023-1905.
51. Richter, J. *et al.* (2024), "EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation", *Proceedings of Interspeech 2024*, pp. 4873–4877, DOI: 10.21437/Interspeech.2024-153.

52. Ji, S., Zuo, J., Fang, M. *et al.* (2024), "TextrolSpeech: A Text Style Control Speech Corpus with Codec Language Text-to-Speech Models", *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10301–10305, DOI: 10.1109/ICASSP48485.2024.10445879.
53. Guan, W., Li, Y., Li, T., Huang, H. *et al.* (2024), "MM-TTS: Multi-modal Prompt based Style Transfer for Expressive Text-to-Speech Synthesis", *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, pp. 18117–18125, DOI: 10.1609/aaai.v38i16.29769.
54. Jin, Z. *et al.* (2024), "SpeechCraft: A Fine-Grained Expressive Speech Dataset with Natural Language Description", *Proceedings of the 32nd ACM International Conference on Multimedia, MM '24*, pp. 1255–1264, DOI: 10.1145/3664647.3681674.
55. Watanabe, A. *et al.* (2023), "Coco-Nut: Corpus of Japanese Utterance and Voice Characteristics Description for Prompt-based Control", *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Taipei, Taiwan, pp. 1–8, DOI: 10.1109/ASRU57964.2023.10389693.
56. Nagrani, A., Chung, J. S., Xie, W. and Zisserman, A. (2020), "Voxceleb: Large-scale speaker verification in the wild", *Computer Speech & Language*, Vol. 60, pp. 101027, DOI: 10.1016/j.csl.2019.101027.
57. Wagner, J., Triantafyllopoulos, A. *et al.* (2023), "Dawn of the transformer era in speech emotion recognition: closing the valence gap", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, no. 9, pp. 10745–10759, DOI: 10.1109/TPAMI.2023.3263585.
58. Sendlmeier, W., Burkhardt, F., Kienast, M., Paeschke, A. and Weiss, B., *The Berlin Database of Emotional Speech*, available at: http://emodb.bilderbar.info/docu/ (Accessed 09.06.25).
59. Meng, R. et al., *SFR-Embedding-Mistral: Enhance Text Retrieval with Transfer Learning*, available at: https://huggingface.co/Salesforce/SFR-Embedding-Mistral (Accessed 4.15.25).
60. Wiseman, J., *Python interface to the WebRTC Voice Activity Detector*, available at: https://github.com/wiseman/py-webrtcvad (accessed 4.15.25).
61. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), *Number Detector and Language Classifier*, available at: https://github.com/snakers4/silero-vad (Accessed 15.04.2025).
62. McFee B. *et al.,* librosa/librosa: 0.11.0. 2025, DOI: 10.5281/zenodo.15006942.
63. Kedrova G. E., Potapov V. V., Egorov A. M. and Omelyanova E. B. (2002), *Akcentologiya. Udarenie i foneticheskoe oformlenie slova. Foneticheskaya priroda udarnyh glasnyh. Russkaya fonetika. Uchebnye materialy.* [Accentology. Word Stress and Phonetic Word Framing. Phonetic Nature of Stressed Vowels. Russian Phonetics. Learning materials.], available at: https://www.philol.msu.ru/~fonetica/akcent/phon_priroda/index.html (Accessed: 09.06.2025).
64. Lubenets I., Davidchuk N. and Amenets A., *Emotions recognition from audio and text files*, available at: https://huggingface.co/datasets/Aniemore/resd_annotated (Accessed 4.15.25).
65. Ardila, R., Branson, M., Davis, K. *et al.* (2020), "Common Voice: A Massively-Multilingual Speech Corpus", *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, pp. 4218–4222, ISBN: 979-10-95546-34-4.

66. Povolockaya, A. A. and Karpov, A. A. (2024), "Analysis of methods for automating extralinguistic component analysis in spontaneous speech. ", *Informatika I Avtomatizaciya*, Vol. 23, no. 1., pp. 5–38, DOI: 10.15622/ia.23.1.1.
67. Lee, Y., Yeon, I., Nam, J. and Chung, J. S. (2023), "VoiceLDM: Text-to-Speech with Environmental Context", DOI: 10.48550/arXiv.2309.13664.
68. Jung, J., Ahn, J., Jung, C., Nguyen, T. D., Jang, Y. and Chung, J. S. (2024), "VoiceDiT: Dual-Condition Diffusion Transformer for Environment-Aware Speech Synthesis", DOI: 10.48550/arXiv.2412.19259.

Информация об авторах:

Е. Н. Радченко – студент 1-го курса магистратуры, Национальный исследовательский технологический университет "МИСИС" (119049, Россия, г. Москва, Ленинский пр-кт, д. 4, стр. 1);

Е. В. Исаева – кандидат филологических наук, доцент, зав. кафедрой английского языка профессиональной коммуникации, Пермский государственный национальный исследовательский университет (614068, Россия, г. Пермь, ул. Букирева, 15), ScopusAuthorID: 57204498718, ResearcherID: O-6777-2015.

Information about the authors:

E. N. Radchenko – 1st year Master's student, National University of Science and Technology "MISIS" (4, B. 1, Leninsky pr. Moscow, Russia, 119049).

E. V. Isaeva – Candidate of Science (in philology), Associate Professor, Head of the Department of English Language of Professional Communication, Perm State University (15, Bukireva St., Perm, Russia, 614068), Scopus Author ID: 57204498718, ResearcherID: O-6777-2015.