Russian Technological Journal.
2025;13(3):7–20

ISSN 2782-3210 (Print)
ISSN 2500-316X (Online)

УДК 004.522
https://doi.org/10.32362/2500-316X-2025-13-3-7-20
EDN PVYBDD

НАУЧНАЯ СТАТЬЯ

Метод конвертации акцента с клонированием голоса в реальном времени на основе неавторегрессионной нейросетевой модели

В.А. Нечаев [@], **С.В. Косяков**

Ивановский государственный энергетический университет имени В.И. Ленина, Иваново, 153003 Россия
[@] *Автор для переписки, e-mail: nechaev@gapps.ispu.ru*

• **Поступила:** 23.07.2024 • **Доработана:** 03.02.2025 • **Принята к опубликованию:** 26.03.2025

Резюме

Цели. В настоящее время при разработке моделей для преобразования речи с акцентом в речь без акцента используются архитектуры глубоких нейросетей, а также ансамбли предобученных нейросетей для распознавания и генерации речи. При этом доступ к реализациям таких моделей является ограниченным, что затрудняет их применение, изучение и дальнейшее развитие. Также использование данных моделей ограничено особенностями архитектуры, которая не позволяет гибко менять тембр генерируемой речи и требует накопления контекста, что ведет к увеличению задержки при генерации и делает данные системы непригодными для использования в сценариях коммуникации двух и более людей в реальном времени. В связи с этим актуальной задачей и целью настоящей работы является разработка метода, позволяющего на основе входной речи с акцентом генерировать речь без акцента с минимальными задержками с возможностью сохранения, клонирования и модификации тембра говорящего, что позволит преодолеть ограничения текущих моделей.

Методы. Применены методы модификации, обучения и объединения глубоких нейросетей в единую сквозную архитектуру для прямого преобразования речи в речь. Для обучения использованы оригинальные и модифицированные наборы данных из открытых источников.

Результаты. Разработан метод конвертации акцента с клонированием голоса в реальном времени на основе неавторегрессионной нейросетевой модели, которая состоит из модулей определения акцента и пола, идентификации говорящего, преобразования речи в фонетическое представление, генерации спектрограммы и декодирования полученной спектрограммы в аудиосигнал. Метод демонстрирует высокое качество конвертации акцента с сохранением оригинального тембра, а также низкие задержки при генерации, приемлемые для использования в сценариях реального времени.

Выводы. Апробация разработанного метода подтвердила эффективность предложенной неавторегрессионной нейросетевой архитектуры. Разработанная прикладная нейросетевая модель продемонстрировала возможность работы в информационных системах на английском языке в режиме реального времени.

Ключевые слова: конвертация акцента, генерация речи, распознавание речи, конвертация голоса, машинное обучение, нейронная сеть

Для цитирования: Нечаев В.А., Косяков С.В. Метод конвертации акцента с клонированием голоса в реальном времени на основе неавторегрессионной нейросетевой модели. *Russian Technological Journal*. 2025;13(3):7–20. https://doi.org/10.32362/2500-316X-2025-13-3-7-20, https://www.elibrary.ru/PVYBDD

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

RESEARCH ARTICLE

Accent conversion method with real-time voice cloning based on a non-autoregressive neural network model

Vladimir A. Nechaev [@], Sergey V. Kosyakov

Ivanovo State Power Engineering University, Ivanovo, 153003 Russia
[@] Corresponding author, e-mail: nechaev@gapps.ispu.ru

• **Submitted:** 23.07.2024 • **Revised:** 03.02.2025 • **Accepted:** 26.03.2025

Abstract

Objectives. The development of contemporary models for the conversion of accents in foreign languages utilizes deep neural network architectures, as well as ensembles of neural networks for speech recognition and generation. However, restricted access to implementations of such models limits their application, study, and further development. Moreover, the use of these models is limited by their architectural features, which prevents flexible changes from being carried out in the timbre of the generated speech and requires the accumulation of context, leading to increased delays in generation, making these systems unsuitable for use in real-time multi-user communication scenarios. Therefore, the relevant task and aim of this work is the development of a method that generates native-sounding speech based on input accented speech material with minimal delays and the capability to preserve, clone, and modify the timbre of the speaker's voice.

Methods. Methods for modifying, training, and combining deep neural networks into a single end-to-end architecture for direct speech-to-speech conversion are applied. For training, original and modified open-source datasets were used.

Results. The work resulted in the development of a real-time accent conversion method with voice cloning based on a non-autoregressive neural network. The model comprises modules for accent and gender detection, speaker identification, speech conversion, spectrogram generation, and decoding the resulting spectrogram into an audio signal. As well as demonstrating high accent conversion quality while maintaining the original timbre, the short generation times of the applied method make it acceptable for use in real-time scenarios.

Conclusions. Testing of the developed method confirmed the effectiveness of the proposed non-autoregressive neural network architecture. The developed model demonstrated the ability to work in real-time information systems in English.

Keywords: accent conversion, speech synthesis, text-to-speech, voice conversion, machine learning, neural network

For citation: Nechaev V.A., Kosyakov S.V. Accent conversion method with real-time voice cloning based on a non-autoregressive neural network model. *Russian Technological Journal.* 2025;13(3):7–20. https://doi.org/10.32362/2500-316X-2025-13-3-7-20, https://www.elibrary.ru/PVYBDD

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

ВВЕДЕНИЕ

Одним из важнейших каналов взаимодействия бизнеса со своими клиентами являются коммуникации посредством голосового общения. Это подтверждается развитием и широким применением колл-центров, которые в настоящее время могут создаваться и работать на межнациональном и межгосударственном уровнях. Для преодоления межъязыкового барьера часто используется английский язык, при том, что операторы и клиенты колл-центров могут не являться носителями этого языка. В результате в практике работы таких колл-центров возникают ситуации, когда клиент отказывается от общения из-за сложности взаимопонимания с оператором, что приводит к экономическим потерям бизнеса. С развитием систем искусственного интеллекта для решения данной проблемы стали применяться программные системы конвертации акцента, которые позволяют в определенной степени уменьшать акцент говорящего [1]. Такие системы могут также применяться в процессе обучения иностранному языку [2, 3], для переозвучивания и улучшения качества ранее записанной речи [4], для улучшения качества распознавания речи существующих систем [5]. Несмотря на то, что в этой области исследований выполнено достаточно большое количество разработок, проблема совершенствования и улучшения качества работы таких систем остается актуальной.

Акцент в речи является неотъемлемой особенностью произношения, его можно разделить на родной, зависящий от множества региональных и культурных факторов, и иностранный [6]. При этом иностранный акцент от родного отличается на сегментном (фонемы) и суперсегментном (интонации, ударения, ритмика) уровнях [7]. Иностранный акцент проявляется, когда носитель одного языка (L1-речь) говорит на другом – неродном или втором для него языке (L2-речь) [8]. L2-речь может быть менее разборчивой для носителей, чем аналогичная по содержанию L1-речь [9], что может привести к снижению уровня понимания и доверия к сказанному, негативному отношению к говорящему и его дискриминации со стороны носителей [10–12].

Ранние методы конвертации акцента на этапе генерации требуют наличия эталонных L1-примеров, соответствующих L2-речи [13–16]. Для каждой L2-фразы необходима соответствующая L1-фраза. Практическое применение таких моделей ограничено ввиду недостаточности данных для покрытия всех возможных речевых вариаций. Такие подходы требуют значительных ресурсов для сбора и обработки данных, что увеличивает время и стоимость разработки таких систем. Кроме того, строгая привязка к парным примерам может снижать универсальность и масштабируемость технологии, ограничивая ее способность адаптироваться к новым акцентам или речевым стилям, которые не были включены в изначальный набор данных.

В следующих разработках это ограничение преодолено – эталонные примеры на этапе вывода не требуются [17–22]. Однако для тренировки моделей [17, 20] используются параллельные наборы данных, содержащие аналогичные L2 и L1-фразы. Получение достаточного объема таких данных является затруднительным и дорогостоящим процессом. Также методы [17, 20] используют авторегрессионные рекуррентные нейросети, что усложняет процесс их тренировки. Метод [19] использует предварительно обученные нейросети для преобразования текста в речь. Для сохранения индивидуальных голосовых характеристик потребуется тренировать отдельную модель для каждого целевого говорящего, что делает затруднительным многопользовательское использование. Методы [21, 22] предсказывают длительность каждой генерируемой фонемы, что в итоге меняет исходную длительность L2-речи и идентичность говорящего, а также требует накопления контекста, что увеличивает время генерации и усложняет использование в реальном времени. Метод [18] лишен недостатков, перечисленных выше, однако реализация модели ограничена конечным набором акцентов, при этом идентификатор акцента должен быть определен на этапе тренировки модели. Это усложняет процесс подготовки данных и применение модели с акцентами, которые не были в них представлены ранее.

Целью данного исследования является разработка метода конвертации акцента, в котором будут преодолены вышеперечисленные проблемы и недостатки. Он должен обеспечивать преобразование любой речи из L2 в L1 без использования эталонных примеров и параллельных данных на этапах обучения

и генерации, что значительно упрощает, удешевляет и ускоряет процесс адаптации системы к новым акцентам.

1. ЗАДАЧИ ИССЛЕДОВАНИЯ

Носители языка воспринимают, с одной стороны, грамотность и беглость речи, с другой – акцент, как отдельные сущности, улучшение одной из них приводит к улучшению общего восприятия L2-речи носителями [9, 23]. При этом абсолютно точное воспроизведение L1-речи неносителем языка на практике сложно достижимо из-за различий в фонетической интерференции и восприятии речи носителями языка [24]. При решении задачи конвертации акцента важно сохранять индивидуальные голосовые особенности говорящего (тембр, высота, громкость), т.е. необходимо производить клонирование голоса [25] при модификации связанных с иностранным акцентом и произношением сегментных и суперсегментных характеристик [13, 17, 18]. Это особенно важно в ситуациях, когда необходимо сохранить эмоциональную окраску, выразительность и индивидуальные особенности речи, особенности голоса, относящиеся к полу говорящего.

Для выполнения этих условий необходимо использовать несколько взаимосвязанных, объединенных в единую сквозную архитектуру моделей для определения акцента и пола, идентификации говорящего (ИГ), преобразования речи в фонетическое представление (РВФ), генерации спектрограммы и декодирования полученной спектрограммы в аудиосигнал.

Также, в сценариях реального времени, например, при голосовом общении людей с использованием высокоскоростных каналов связи, необходимо обеспечить минимальные задержки на генерацию и передачу обработанной речи [26–28]. Перевод акцента должен выполняться в режиме реального времени без использования рекуррентных сетей [29], чтобы избежать эффекта накопления ошибок, связанного с последовательной генерацией вывода.

На этапе тренировки должны использоваться общедоступные открытые данные для обучения систем распознавания и генерации речи. Также метод должен быть независим от наличия эталонных примеров и параллельных данных на этапах обучения и вывода.

2. МЕТОДЫ ИССЛЕДОВАНИЯ

2.1. Архитектура разработанной системы

Разработанный метод конвертации акцента включает в себя несколько взаимосвязанных, объединенных в единую сквозную архитектуру моделей для определения акцента и пола, ИГ, преобразования РВФ, генерации спектрограммы и декодирования полученной спектрограммы в аудиосигнал. На рис. 1 представлена общая схема взаимодействия указанных моделей на этапе вывода (генерации выходного L1-аудио).

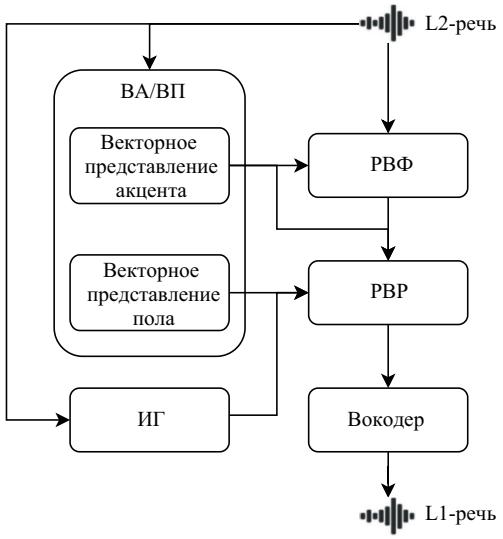

Рис. 1. Общая схема вывода метода конвертации акцента с клонированием голоса

Аудиосигнал L2-речи подается на вход модели РВФ, на вход модели определения и векторизации акцента и пола (ВА/ВП) и на вход модели ИГ. Векторное представление акцента влияет на генерацию фонетического представления, которое в векторной форме подается на вход модели преобразования речи в речь (РВР) и генерации мел-спектрограммы. Также на вход РВР-модели подаются векторные представления акцента (ВА) и пола (ВП), а также выход ИГ-модели, представляющий собой векторное представление индивидуальных голосовых характеристик (тембра). Полученная спектрограмма преобразуется в аудиосигнал L1-речи с помощью декодирующей модели вокодера.

Общий конвейер генерации L1-речи из исходной L2-речи может быть упрощенно представлен в виде формулы:

$$a_{L1} = F_V(F_{РВР}(F_{РВФ}(a_{L2}, F_{ВА}(a_{L2})), F_{ВА}(a_{L2}), F_{ВП}(a_{L2}), F_{ИГ}(a_{L2}))), \quad (1)$$

где a_{L1} – сгенерированный аудиосигнал L1-речи; a_{L2} – входной аудиосигнал L2-речи; F_V – модель вокодера; $F_{РВР}$ – РВР-модель; $F_{РВФ}$ – РВФ-модель; $F_{ВА}$ – ВА/ВП-модель, ВА; $F_{ВП}$ – ВА/ВП-модель, ВП; $F_{ИГ}$ – ИГ-модель, векторное представление индивидуальных голосовых характеристик.

Для получения единой сквозной модели конвертации акцента необходимо последовательно выполнять процесс тренировки каждой модели.

Так, ВА/ВП- и ИГ-модели не зависят от других моделей и их тренировку можно проводить в любом порядке. На этапе получения РВФ-модели потребуется вывод готовой ВА/ВП-модели. Для получения РВР-модели необходимы все предыдущие модели (ВА/ВП, ИГ, РВФ). Для тренировки модели вокодера требуется вывод РВР-модели.

2.2. ВА/ВП-модель

Для получения векторов фиксированной длины, представляющих свойства акцента и пола говорящего, модель сначала обучалась для решения задачи классификации. В такой конфигурации в процессе обучения используются метки классов, которые модель возвращает на выходе последнего слоя, а векторные представления, используемые в качестве голосовых характеристик, берутся со специального промежуточного слоя.

В данной и остальных моделях применяется препроцессор на основе быстрого преобразования Фурье, который переводит входящий аудиосигнал (временная область) в мел-спектрограмму (частотная область), показывающую частотное содержание аудиосигнала на перцептивной мел-шкале, которая аппроксимирует нелинейную частотную характеристику человеческого уха. Частота дискретизации (семплирования) – 22050 Гц, ширина окна – 1024 звуковых фрагмента (семпла), шаг окна – 256 семплов, количество генерируемых мел-диапазонов – 80.

На рис. 2 представлена схема тренировки модели определения акцента и пола. Она содержит блоки сверточной сети архитектуры Jasper конфигурации 3 × 3 [30]. Декодер акцента и декодер пола имеют одинаковую архитектуру и состоят из слоя объединения (attention pooling layer) [31], слоя нормализации, сверточного слоя для получения ВА и ВП размерности 192, а также линейного слоя для получения (предсказания) класса акцента (КА) и пола (КП).

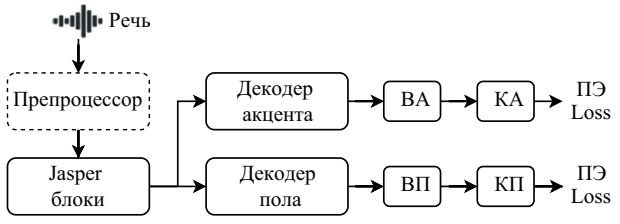

Рис. 2. Схема тренировки ВА/ВП-модели. ПЭ – перекрестные энтропии

Аудиосигнал поступает в препроцессор, далее мел-спектрограмма поступает в Jasper блоки и, параллельно, в декодер акцента и декодер пола, а также соответствующие полносвязные слои на выходе для получения векторов предсказаний акцента и пола. В процессе обучения модели минимизируется сумма ПЭ:

$$L_{\text{ВА, ВП}} = (x_{\text{a}}, y_{\text{a}}, x_{\text{g}}, y_{\text{g}}) = \\ = -\sum_{i=1}^{A} y_{\text{a}_i} \ln\left(\frac{\exp x_{\text{a}_i}}{\sum_{k=1}^{A} \exp x_{\text{a}_k}}\right) - \sum_{j=1}^{G} y_{\text{g}_j} \ln\left(\frac{\exp x_{\text{g}_j}}{\sum_{l=1}^{G} \exp x_{\text{g}_l}}\right), \quad (2)$$

где $L_{\text{ВА, ВП}}$ – общая функция потерь ВА/ВП-модели, A – количество КА (40), G – количество КП (2), x_{a} – предсказания акцента, x_{g} – предсказания пола, y_{a} – реальные метки акцента, y_{g} – реальные метки пола.

2.3. Модель ИГ

На рис. 3 представлена схема тренировки ИГ-модели и векторного представления тембра. Она содержит входящую сверточную нейросеть архитектуры SincNet [32], слои модели X-Vectors DNN[1] [33], слой для получения векторных представлений размерности 512.

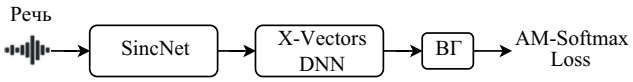

Рис. 3. Схема тренировки ИГ-модели. ВГ – векторное представление индивидуальных голосовых характеристик говорящего

В отличие от ВА/ВП-модели, предварительное преобразование аудиосигнала в мел-спектрограмму не производится, т.е. оцифрованный звуковой сигнал во временной области с частотой дискретизации 16000 Гц поступает на полосовые фильтры архитектуры SincNet, далее – на сверточные слои X-Vectors DNN и выходной полносвязный слой, дающий на выходе векторное представление индивидуальных голосовых характеристик (тембра). В процессе тренировки модели решается задача обучения представлениям (representation learning) и минимизируется функция потерь Additive Angular Margin (AAM Loss) [34].

2.4. Модель преобразования РВФ

Следующим этапом является распознавание речи с учетом акцента говорящего. Для этого необходимо получить модель преобразования речи в фонетическое или текстовое представление. Схема тренировки РВФ-модели показана на рис. 4. На ней пунктирной линией обозначены блоки, которые фиксируются в процессе обратного распространения ошибки, т.е. веса в данных блоках не обновляются, а используются их заранее полученные состояния.

[1] Deep neural network.

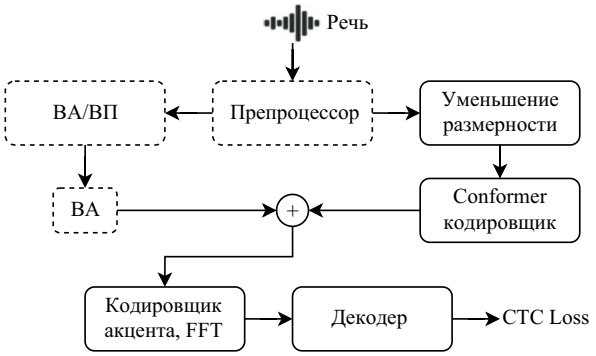

Рис. 4. Схема тренировки РВФ-модели

Речевой аудиосигнал поступает на вход препроцессора, описанного ранее, и далее параллельно в ВА/ВП-модель для получения ВА и в блок сверточного уменьшения размерности (Subsampler) с фактором 4 с дальнейшим преобразованием с помощью Conformer кодировщика, который представляет собой 12 модулей архитектуры Conformer [35] с внутренней размерностью 512, состоящих из полносвязных [36], сверточных [37] и трансформерных слоев с механизмом внимания [38]. Затем ВА нормализуется, приводится к размерности 512, суммируется с выходным сигналом Conformer кодировщика и поступает на вход кодировщика акцента, который имеет архитектуру одного стека трансформера прямой связи (feed-forward transformer, FFT) [39]. Выходной сигнал кодировщика акцента в дальнейшем используется в РВР-модели, как распределение фонетических токенов. В заключении выходной сигнал кодировщика акцента поступает на декодер, имеющий однослойную сверточную архитектуру с многопеременной логистической функцией активации (Softmax), формирующий на выходе вектор предсказаний текстовых токенов размерности, равной размеру словаря токенизатора (128) плюс один (для пустого «blank» токена). В процессе обучения модели минимизируется функция, которая вычисляет потери между непрерывным (несегментированным) временным рядом и целевой последовательностью (CTC Loss) [40]:

$$L_{\text{РВФ}}(x, y) = -\ln\left(\sum_{\rho \in A_{x,y}} \prod_{t=1}^{T} x_{\rho_t}\right), \qquad (3)$$

где $L_{\text{РВФ}}$ – функция потерь РВФ-модели (Connectionist Temporal Classification (CTC) Loss), x – предсказанные моделью вероятности текстовых токенов, y – последовательность текстовых токенов из целевого текста, ρ – путь выравнивания x предсказаний для сокращения до y последовательности путем удаления всех пустых («blank») токенов и слияния повторяющихся токенов, $A_{x,y}$ – множество всех возможных путей выравнивания, T – количество предсказанных токенов в x, x_{ρ_t} – вероятность конкретного предсказанного токена на шаге t при выбранном пути выравнивания ρ.

2.5. Модель преобразования РВР и генерации спектрограммы

Предыдущие модели объединяются в единую архитектуру для преобразования РВР и генерации спектрограммы. На рис. 5 представлена схема ее тренировки. В состав РВР-модели входят рассмотренные ранее блоки препроцессора, определения акцента, пола (ВА/ВП-модель) и тембра говорящего (ИГ-модель) с соответствующими модулями векторных представлений (ВА, ВП, ВГ), а также блок преобразования РВФ (РВФ-модель). Все эти блоки обозначены пунктирной линией – их обучение было выполнено ранее и не проводится на этапе тренировки РВР-модели. Также в архитектуру добавлен необучаемый блок на основе нормализованной взаимно корреляционной функции (normalized cross correlation function) и медианного сглаживания для извлечения основной или самой низкой частоты периодического звукового сигнала (F0), которая воспринимается человеческим ухом как высота тона [41, 42].

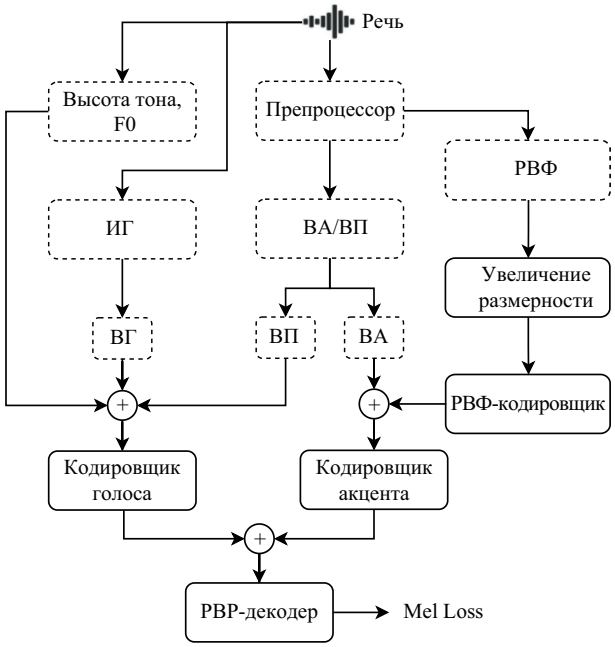

Рис. 5. Схема тренировки РВР-модели

Речевой аудиосигнал поступает на препроцессор, блок высоты тона и вход ИГ-модели. Мел-спектрограмма с препроцессора подается на входы ВА/ВП- и РВФ-моделей. Фонетическое представление из РВФ-модели поступает в блок увеличения размерности (upsampler) с фактором 4, чтобы выровнять исходную и генерируемые спектрограммы, состоящий из двух сверточных

1D-транспонированных слоев и двух функций активации выпрямителя (rectified linear unit, ReLU), располагающихся после каждого сверточного слоя. После блока увеличения размерности фонетическое представление преобразуется с помощью РВФ-кодировщика, который имеет архитектуру шести стеков трансформера прямой связи (FFT) [39], применяющегося в архитектуре Fastpitch в качестве входного блока, работающего в области токенов [43], с внутренней и внешней размерностями, соответственно, 1536 и 384. Векторные представления акцента, пола, тембра говорящего и профиля высоты тона нормализуются и приводятся к размерности 384. Далее векторы акцента и выход РВФ кодировщика суммируются и подаются на вход кодировщика акцента (1 стек FFT). Аналогично суммируются векторы высоты тона, тембра, пола и поступают на вход кодировщика голоса (1 стек FFT). Таким образом, кодировщик голоса агрегирует свойства речи, относящиеся к индивидуальным голосовым характеристикам кроме акцента, за который, в свою очередь, отвечает кодировщик акцента. Сумма выходных векторов кодировщика голоса и кодировщика акцента поступает на вход РВР-декодера, состоящего из 6 стеков FFT архитектуры Fastpitch из выходной мел-области [43]. В завершении, вектор проецируется в размерность 80 для соответствия исходному количеству мел-диапазонов. В процессе обучения минимизируется функция потерь на основе среднеквадратической ошибки:

$$L_{\text{PBP}}(x,y)=\frac{1}{\sum_{i=1}^{N} d_i}\sum_{i=1}^{N} d_i (y_i - x_i)^2, \qquad (4)$$

где L_{PBP} – функция потерь РВР-модели (Mel Loss), N – количество элементов в мел-спектрограмме, x – предсказанная моделью мел-спектрограмма, y – целевая мел-спектрограмма, d – маска длительности спектрограммы для сбора в партию (batch) фиксированного размера, состоящая из значений 1 («элемент нужно учитывать») и 0 («элемент учитывать не нужно»), полученная из длительности предсказанной спектрограммы.

2.6. Модель генерации звукового сигнала из мел-спектрограммы (вокодер)

Мел-спектрограмма L1-речи в частотной области, полученная с помощью РВР-модели, преобразуется в звуковой сигнал во временной области. Для этого используется модель на основе генеративно-состязательных сетей HiFi-GAN [44]. Аудиосигнал на выходе имеет частоту дискретизации 22050 Гц. Обучение модели проводится следующим образом: аудиосигнал из обучающего набора данных конвертируется в мел-спектрограмму с помощью РВР-модели, далее полученная спектрограмма передается в вокодер и преобразуется в звуковой сигнал. С помощью полученного и исходного аудиосигналов рассчитываются функции потерь для генератора и дискриминатора, описанные в [44].

2.7. Упрощенная модель преобразования РВР и генерации спектрограммы (Ablation)

С целью проведения сравнительных экспериментов также была разработана упрощенная версия модели конвертации акцента, схема которой представлена на рис. 6.

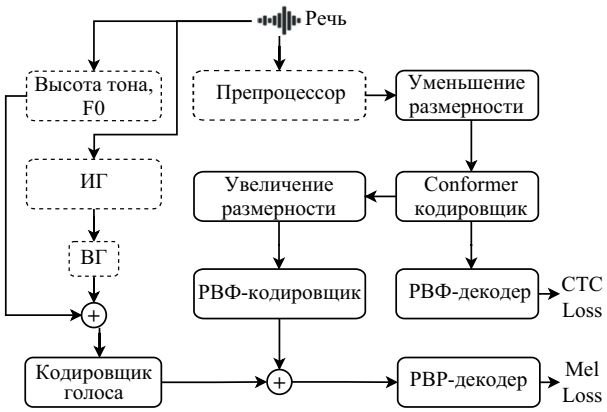

Рис. 6. Схема упрощенной модели конвертации акцента

Из данной упрощенной модели исключена ВА/ВП-модель, а также все связанные с ней кодировщики в РВФ- и РВР-моделях. Таким образом, в полученной упрощенной модели выходной сигнал не обуславливается свойствами акцента и пола. Кроме этого, обучение РВФ-модели проводилось не отдельно, а одновременно с РВР-моделью без фиксации весов РВФ с минимизацией суммы функций потерь CTC Loss и Mel Loss.

3. ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ МЕТОДА

3.1. Обучение моделей

Обучение ВА/ВП-модели проводилось на следующих наборах данных: CMU-ARCTIC [45], L2 ARCTIC [46], Speech Accent Archive [47], Common Voice 16.1 [48]. Все они представляют собой аудиозаписи речи на английском языке, соответствующие им текстовые транскрипции, а также содержат дополнительную метаинформацию об акценте, поле и, в некоторых случаях, о родном языке, месте проживания и возрасте говорящего. Используя эту информацию, аудиофайлы были сгруппированы

в 40 классов, обозначающих родной или иностранный английский акцент, например, британский, американский, русский, индийский и южноазиатский, канадский, немецкий, австралийский, африканский, японский, восточноевропейский и т.д. Также выделен пол говорящего. Общая длительность размеченных таким образом аудиофайлов составила для тренировочной выборки – 1087.6 ч, для валидационной и тестовой – 7.6 ч.

Для обучения ИГ-модели использованы коллекции данных VoxCeleb1 [49] и VoxCeleb2 [50] общей продолжительностью 2794 ч. Эти наборы представляют собой сгруппированные аудиозаписи речи 7363 людей. Аудиозаписи, относящиеся к одному человеку, подаются при обучении как положительные примеры и, наоборот, относящиеся к разным людям, как негативные примеры.

Модель РВФ тренировалась на данных CMU-ARCTIC [45], L2 ARCTIC [46], Common Voice 16.1 [48], LibriSpeech [51], NPTEL2020[2], VCTK [52], GigaSpeech [53]. Указанные наборы состоят из аудиозаписей англоязычной речи с разными акцентами и соответствующих текстовых транскрипций. Общая длительность объединенной тренировочной выборки – 6107 ч, валидационной – 48 ч. Текстовые транскрипции были нормализованы, т.е. преобразованы из канонической письменной формы в устную [54], что особенно важно для чисел и аббревиатур, а также были приведены к единому виду: переведены в нижний регистр в ASCII-формат, удалена пунктуация, специальные символы и дополнительные отступы. На тренировочной части текстов был обучен токенизатор SentencePiece [55] с размером словаря 128, с помощью которого в процессе тренировки и оценки модели обрабатываются все тексты.

Обучение РВР-модели и вокодера производилось с использованием следующих наборов данных: CMU-ARCTIC [45], L2 ARCTIC [46], VCTK [52], LibriTTS-R [56], LJ Speech[3]. При разделении данных на тренировочную и валидационную выборки их длительность составила 681 ч и 17.6 ч соответственно. В процессе тренировки используется только аудиоинформация без текстовой разметки.

Обучение упрощенной модели (Ablation) проводилось на данных для РВФ- и РВР-моделей.

Для обучения, оценки и использования описанных моделей разработан код с использованием библиотек с открытым исходным кодом Pytorch [57] и NVIDIA NeMo [58]. Реализация и веса модели векторного представления тембра говорящего (ИГ-модель) взяты из библиотеки Pyannote [59]. Обучение проводилось на сервере с 8 графическими ускорителями (graphics processing unit, GPU) NVIDIA Tesla V100.

Обучение ВА/ВП-модели проводилось с использованием оптимизатора SGD со скоростью обучения (learning rate) $1 \cdot 10^{-3}$, коэффициентом регуляризации (weight decay) $2 \cdot 10^{-4}$, коэффициентом инерции (momentum) 0.9 и планировщиком Cosine Annealing в течение 200 эпох. Для тренировки моделей РВФ и РВР применялся оптимизатор AdamW при скорости обучения $1 \cdot 10^{-3}$, коэффициенте регуляризации 0.001 и аналогичном планировщике, что и для ВА/ВП-модели, в течение 50 эпох для каждой модели. Тонкая настройка модели вокодера HiFi-GAN осуществлялась с инициализации весов модели, полученных из открытых источников [44], с использованием оптимизатора AdamW и скорости обучения $1 \cdot 10^{-6}$ в течение 40 эпох. Обучение упрощенной модели (Ablation) проводилось с аналогичными параметрами, применяемым в РВР-модели.

В табл. 1 представлено количество тренируемых параметров моделей, оптимизируемых в процессе обучения. Всего рассмотренная архитектура преобразования акцента (полная РВР), состоящая из нескольких взаимосвязанных моделей, имеет 164 млн параметров.

Таблица 1. Количество тренируемых параметров

Модель	Количество параметров, млн
ВА/ВП	24.9
ИГ	4.3
РВФ	82.1
РВР	52.7
Полная РВР	**164**
Вокодер	84.7
Всего	248.7

3.2. Оценка производительности

Оценка производительности модели осуществлялась на сервере под управлением операционной системы Linux с одним графическим вычислителем (GPU) NVIDIA Tesla T4, виртуальным процессором (virtual central processing unit, vCPU) с 8 ядрами и 16 ГБ оперативной памяти (random access memory, RAM). Для этого модель сначала была экспортирована в открытый формат ONNX и затем развернута с использованием программного обеспечения с открытым исходным кодом NVIDIA Triton.

[2] NPTEL2020 – Indian English Speech Dataset. https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset. Дата обращения 01.05.2024. / Accessed May 01, 2024.

[3] Ito K., Johnson L. *The LJ Speech Dataset*. https://keithito.com/LJ-Speech-Dataset/. Дата обращения 01.05.2024. / Accessed May 01, 2024.

Используя программный интерфейс развернутой в NVIDIA Triton-модели и тестовый аудиофайл продолжительностью 5 с, содержащий англоязычную L2-речь, были выполнены замеры задержки генерации ответа при 200 итерациях. В итоге средняя задержка при генерации (latency) составила 52 мс, пропускная способность – 96 RTFX.

Результаты оценки производительности модели конвертации акцента показывают низкие задержки при генерации. Вместе с особенностями архитектуры, которая не требует накопления длинного контекста, а может работать с отрезками длительностью менее 0.25 с, это дает возможность применения предлагаемой модели в реальном времени в режиме диалога, когда задержки с ответом влияют на коммуникацию [26–28].

3.3. Объективная оценка качества

Для проведения объективной оценки качества были использованы данные из открытых источников, а также предварительно обученные модели распознавания речи. С помощью предложенного метода конвертации акцента для каждого примера из тестового набора был сгенерирован аудиофайл. Далее рассчитывались метрики качества для оригинального и исправленного аудиофайлов. В табл. 2. представлены результаты объективной оценки качества.

В качестве тестовых наборов данных использованы подвыборки общей продолжительностью 26.9 ч, которые не участвовали в процессе тренировки модели конвертации акцента и ее составляющих. Все они включают в себя текстовые транскрипции и аудиофайлы с речью на английском языке с разными родными и неродными акцентами из открытых источников:

- 3.2 ч из CMU-ARCTIC [45], L2 ARCTIC [46] (ARCTIC), 10 акцентов: американский, английский, китайский, индийский, корейский, вьетнамский, испанский, арабский, голландский, немецкий;
- 3.1 ч из Common Voice [48], 12 акцентов: американский, английский, индийский, австралийский, африканский, китайский, филиппинский, малазийский, немецкий, русский, французский, восточноевропейский;
- 15.2 ч из NPTEL2020, индийский акцент;
- 5.4 ч из Afrispeech-200 [60], африканский акцент (йоруба, суахили, игбо, зулу, тсвана, идома, африкаанс).

Были использованы модели распознавания речи, полученные из открытых источников: Conformer [35], Citrinet [61], Whisper [62]. При этом модель Whisper взята в двух вариантах: «большая» мультиязычная (L. Mult.) и «средняя» англоязычная (M. En.). Распознавание проводилось на аудиофайлах без обработки и на аудиофайлах после конвертации акцента. Затем распознанные и истинные транскрипции приводились к единому виду с помощью нормализации [54], после чего сравнивались и считались метрики качества: частота ошибок в словах (word error rate, WER), частота ошибок в символах (character error rate, CER). В таблице жирным шрифтом выделены наилучшие результаты для каждой пары: тестовый набор данных и модель распознавания речи.

Таблица 2. Результаты оценки модели конвертации акцента с помощью моделей распознавания речи. Данные после конвертации отмечены, как «конв.»

Тестовый набор данных	Модель распознавания речи			
	Conformer	Citrinet	Whisper L. Mult.	Whisper M. En.
WER, %				
ARCTIC	9.57	11.73	16.23	8.91
ARCTIC конв.	**8.78**	**11.55**	**12.69**	**8.68**
Common Voice	**9.07**	25.80	36.89	11.26
Common Voice конв.	9.12	**23.38**	**22.71**	**10.62**
NPTEL2020	29.18	29.88	16.41	15.18
NPTEL2020 конв.	**25.26**	**29.41**	**13.87**	**11.64**
Afrispeech-200	43.2	46.24	37.91	33.61
Afrispeech-200 конв.	**35.19**	**39.49**	**35.56**	**29.96**
CER, %				
ARCTIC	3.73	4.85	10.30	3.98
ARCTIC конв.	**3.52**	**4.68**	**6.06**	**3.92**
Common Voice	**3.75**	8.74	21.41	5.66
Common Voice конв.	3.77	**8.29**	**13.63**	**5.22**
NPTEL2020	16.87	17.70	11.94	10.67
NPTEL2020 конв.	**14.79**	**17.01**	**10.10**	**9.44**
Afrispeech-200	31.52	34.79	24.30	20.04
Afrispeech-200 конв.	**27.86**	**28.92**	**23.15**	**18.88**

Как видно из результатов, практически во всех случаях применение метода конвертации акцента улучшает распознавание предварительно обученных моделей, на что указывают пониженные значения частот ошибок в словах и символах. Модель конвертации акцента улучшает качество речи, делая ее более распознаваемой.

3.4. Субъективная оценка качества

Были проведены тесты на звуковое восприятие, основанные на мнении группы людей, в которых приняли участие 53 человека из разных стран

с уровнем владения английским языком не ниже B2 согласно шкале CEFR[4]. Для этого каждому из участников были даны инструкции, где в рамках каждого эксперимента предлагалось прослушать 1 или 2 аудиофайла и дать свою оценку соответствия критерию качества по пятибалльной шкале, где «1» – точно не соответствует, «2» – скорее не соответствует, «3» – компромисс, «4» – скорее соответствует, «5» – точно соответствует. Далее полученные оценки были использованы для расчета средней экспертной оценки (mean opinion score, MOS) для каждого эксперимента. Результаты представлены в табл. 3.

В качестве звуковых образцов были случайно отобраны 20 пар аудиофайлов из тестовых подвыборок наборов данных L2 ARCTIC [46] и NPTEL2020 с неродным английским акцентом (Original): индийский, китайский, корейский, вьетнамский, испанский, арабский, немецкий. Каждая оригинальная аудиопара представляет собой запись одного и того же говорящего. Для каждого отобранного аудиофайла (всего 40) были сгенерированы варианты с использованием упрощенной модели конвертации акцента (Ablation) и с помощью предлагаемой модели (Proposed). Всего проведено 3 эксперимента для оценки натуральности голоса, сходства говорящих и отсутствия иностранного акцента. Во всех экспериментах предлагалось сделать не менее 3 оценок для каждого типа звукового образца. При этом сами тестовые образцы менялись в экспериментах, исключая повторение. Перед выставлением оценки можно было прослушать образец неограниченное число раз. Таким образом, каждый опрошенный сделал в общей сложности от 9 до 12 оценок.

При оценке натуральности голоса участникам предлагалось определить по пятибалльной шкале, насколько «натурально» звучит речь в аудиопримере, т.е. создается ли впечатление у слушателя, что это настоящий живой человеческий голос, а не сгенерированная или роботизированная речь. Оценка «1» означает, что голос точно искусственный, синтезированный с использованием методов компьютерной генерации, а «5» – что в примере звучит речь, полученная с использованием способов аналоговой или цифровой звукозаписи голоса реального человека. Также опрошенным были даны рекомендации не обращать внимание на наличие или отсутствие фонового шума в записи, чтобы сконцентрироваться именно на оценке речи.

[4] Шкала CEFR (Common European Framework of Reference) – система уровней владения иностранным языком, используемая в Европе. https://www.coe.int/en/web/common-european-framework-reference-languages. Дата обращения 01.05.2024. [CEFR (Common European Framework of Reference) is the system of foreign language proficiency levels used in Europe. https://www.coe.int/en/web/common-european-framework-reference-languages. Accessed May 01, 2024.]

Таблица 3. Результаты субъективной оценки качества (MOS с 95% доверительным интервалом)

Примеры	Натуральность голоса	Сходство говорящих	Отсутствие иностранного акцента
Original	4.83 ± 0.10	4.91 ± 0.08	2.06 ± 0.18
Ablation	3.38 ± 0.13	3.92 ± 0.15	3.58 ± 0.17
Proposed	4.04 ± 0.16	4.30 ± 0.18	4.11 ± 0.14

С целью проведения эксперимента по оценке сходства говорящих были подготовлены пары аудиозаписей: Original – Original, Original – Ablation и Original – Proposed. При этом в первую пару входят записи только из оригинальных данных, представляющих собой записи одного и того же говорящего, но произносящего разные фразы. В другие пары входит оригинальная запись одной фразы и сгенерированный вариант другой фразы того же самого говорящего. Участникам предлагалось прослушать такие пары аудиозаписей и решить, произнесены ли они одним и тем же человеком, т.е. насколько тембр в одном файле похож на тембр в другом файле. Оценка «1» – речь в аудиозаписях точно принадлежит разным людям, «5» – тембр говорящих в аудиозаписях идентичный, принадлежащий одному человеку. Опрошенным рекомендовано во время оценки не обращать внимание на свойства L1 и L2-акцента, чтобы сосредоточиться на сравнении обертоновой окраски голоса.

Для оценки отсутствия иностранного акцента участникам предлагалось прослушать англоязычный аудиофайл и решить, на сколько, по их мнению, в данной записи выражен иностранный акцент. Английский и американский акценты приняли считать родными L1, а все остальные – неродными L2. Оценка «1» означает, что в речи ярко выражен иностранный L2-акцент, «5» – речь точно является англоязычной L1 без иностранного акцента.

Анализ таблицы показывает, что наивысшие оценки натуральности голоса и сходства говорящих показывают оригинальные примеры, что очевидно, т.к. они получены без использования методов синтеза речи, а вместе с наименьшей оценкой отсутствия иностранного акцента демонстрирует калибровку мнений участников эксперимента на реальных данных. Добавление BA/BП-модели в общую схему модели конвертации акцента существенно повышает качество генерации, это демонстрируют повышенные результаты по сравнению с упрощенной Ablation-моделью. Во всех субъективных экспериментах предложенная модель показывает оценку выше «4», означающую, что, по мнению опрошенных, модель скорее соответствует обозначенным критериям качества.

ЗАКЛЮЧЕНИЕ

В работе представлен метод конвертации акцента, который позволяет преобразовывать любую L2-речь с выраженным иностранным акцентом в L1-речь, не зависит от наличия эталонных примеров и параллельных данных на этапах обучения и генерации, что значительно упрощает, удешевляет и ускоряет процесс адаптации системы к новым акцентам.

Предложенная модель является неавторегрессионной и не использует в своей архитектуре рекуррентные сети, что позволяет ускорить процесс тренировки, выполнять перевод акцента в режиме реального времени, избежать эффекта накопления ошибок, связанного с последовательной генерацией вывода.

Рассматриваемый метод также включает алгоритм клонирования речевых характеристик голоса, благодаря которому сохраняется идентичность говорящего даже после преобразования акцента. Это особенно важно в ситуациях, когда необходимо сохранить эмоциональную окраску, выразительность и индивидуальные особенности речи. Кроме того, метод позволяет в реальном времени в процессе генерации модифицировать характеристики голоса, такие как акцент, тембр и особенности голоса, относящиеся к полу говорящего, путем копирования соответствующих характеристик из аудиообразца, что делает его применимым в более широком круге сценариев, чем прежние разработки.

Модель демонстрирует высокое качество конвертации акцента с сохранением оригинального тембра, а также низкие задержки при генерации, приемлемые для использования в сценариях реального времени.

Применение метода позволяет:

1) преобразовывать англоязычную речь с иностранным L2-акцентом в L1-речь без иностранного акцента;
2) улучшать качество речи и, как следствие, повышать качество распознавания существующих систем;
3) копировать и менять голосовые характеристики говорящего в реальном времени;
4) применять конвертацию акцента в реальном времени в режиме диалога.

Разработанная прикладная нейросетевая модель продемонстрировала возможность работы в информационных системах на английском языке в режиме реального времени. Результаты исследования могут применяться при разработке систем модификации голоса, а также распознавания и генерации речи.

Вклад авторов

Все авторы в равной степени внесли свой вклад в исследовательскую работу.

Authors' contribution

All authors equally contributed to the research work.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. McMillin D.C. Outsourcing identities: Call centres and cultural transformation in India. *Economic and Political Weekly*. 2006;41(3):235–241.
2. Felps D., Bortfeld H., Gutierrez-Osuna R. Foreign accent conversion in computer assisted pronunciation training. *Speech Communication*. 2009;51(10):920–932. https://doi.org/10.1016/j.specom.2008.11.004
3. Probst K., Ke Y., Eskenazi M. Enhancing foreign language tutors–In search of the golden speaker. *Speech Communication*. 2002;37(3–4):161–173. https://doi.org/10.1016/S0167-6393(01)00009-7
4. Türk O., Arslan L. M. Subband based voice conversion. In: *7th International Conference on Spoken Language Processing, ICSLP2002 – INTERSPEECH 2002*. *Interspeech*. 2002. P. 289–292.
5. Biadsy F., Weis, R.J., Moreno P.J., Kanevsky D., Jia Y. Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. *Interspeech*. 2019. P. 4115–4119. http://doi.org/10.21437/Interspeech.2019-1789
6. Birner B. *Why Do Some People Have an Accent?* Linguistic Society of America. Washington, DC. 1999. 6 p.
7. Baese-Berk M.M., Morrill T.H. Speaking rate consistency in native and non-native speakers of English. *J. Acoust. Soc. Am*. 2015;138(3):EL223–EL228. https://doi.org/10.1121/1.4929622
8. Piske T., MacKay I.R.A., Flege J.E. Factors affecting degree of foreign accent in an L2: A review. *J. Phonetics.* 2001;29(2):191–215. https://doi.org/10.1006/jpho.2001.0134
9. Munro M.J., Derwing T.M. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*. 1995;45(1):73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x
10. Lev-Ari S., Keysar B. Why don't we believe non-native speakers? The influence of accent on credibility. *J. Exp. Soc. Psychol*. 2010;46(6):1093–1096. https://doi.org/10.1016/j.jesp.2010.05.025
11. Rubin D.L., Smith K.A. Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *Int. J. Intercult. Relat*. 1990;14(3):337–353. https://doi.org/10.1016/0147-1767(90)90019-S

12. Nelson Jr. L.R., Signorella M.L., Botti K.G. Accent, gender, and perceived competence. *Hispanic J. Behavior. Sci.* 2016;38(2):166–185. https://doi.org/10.1177/0739986316632319
13. Zhao G., Gutierrez-Osuna R. Using phonetic posteriorgram based frame pairing for segmental accent conversion. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2019;27(10):1649–1660. https://doi.org/10.1109/TASLP.2019.2926754
14. Zhao G., Sonsaat S., Levis J., Chukharev-Hudilainen E., Gutierrez-Osuna R. Accent conversion using phonetic posteriorgrams. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2018. P. 5314–5318. https://doi.org/10.1109/ICASSP.2018.8462258
15. Aryal S., Gutierrez-Osuna R. Can voice conversion be used to reduce non-native accents? In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2014. P. 7879–7883. https://doi.org/10.1109/ICASSP.2014.6855134
16. Ding S., Zhao G., Gutierrez-Osuna R. Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning. *Computer Speech & Language*. 2022;72:101302. https://doi.org/10.1016/j.csl.2021.101302
17. Quamer W., Das A., Levis J., Chukharev-Hudilainen E., Gutierrez-Osuna R. Zero-shot foreign accent conversion without a native reference. *Proc. Interspeech*. 2022. http://doi.org/10.21437/Interspeech.2022-10664
18. Jin M., Serai P., Wu J., Tjandra A., Manohar V., He Q. Voice-preserving zero-shot multiple accent conversion. In: *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2023. P. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094737
19. Zhou Y., Wu Z., Zhang M., Tian X., Li H. TTS-guided training for accent conversion without parallel data. *IEEE Signal Proc. Lett*. 2023;30:533–537. https://doi.org/10.1109/lsp.2023.3270079
20. Zhao G., Ding S., Gutierrez-Osuna R. Converting foreign accent speech without a reference. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2021;29:2367–2381. https://doi.org/10.1109/TASLP.2021.3060813
21. Liu S., Wang D., Cao Y., Sun L., Wu X., Kang S., Wu Z., Liu X., Su D., Yu D., Meng H. End-to-end accent conversion without using native utterances. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2020. P. 6289–6293.
22. Zhou X., Zhang M., Zhou Y., Wu Z., Li H. Accented text-to-speech synthesis with limited data. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2024;32:1699–1711. https://doi.org/10.1109/TASLP.2024.3363414
23. Pinget A.F., Bosker H.R., Quené H., De Jong, N.H. Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*. 2014;31(3):349–365. https://doi.org/10.1177/0265532214526177
24. Бархударова Е.Л. Методологические проблемы анализа иностранного акцента в русской речи. *Вестник Московского университета. Серия 9. Филология.* 2012;6:57–70.
[Barkhudarova E.L. Methodological Problems in Analyzing Foreign Accents in Russian Speech. *Vestnik Moskovskogo universiteta. Seriya 9. Filologiya = Lomonosov Philology J.* 2012;6:57–70 (in Russ.).]
25. Arik S., Chen J., Peng K., Ping W., Zhou Y. Neural voice cloning with a few samples. *Advances in Neural Information Processing Systems* (*NeurIPS 2018*). 2018;31.
26. Cohen D. Issues in transnet packetized voice communication. In: *Proceedings of the fifth Symposium on Data Communications* (*SIGCOMM'77*). 1977. P. 6.10–6.13. https://doi.org/10.1145/800103.803349
27. Liang Y.J., Farber N., Girod B. Adaptive playout scheduling and loss concealment for voice communication over IP networks. *IEEE Trans. Multimedia*. 2003;5(4):532–543. https://doi.org/10.1109/TMM.2003.819095
28. Matzinger T., Pleyer M., Żywiczyński P. Pause Length and Differences in Cognitive State Attribution in Native and Non-Native Speakers. *Languages*. 2023;8(1):26. http://doi.org/10.3390/languages8010026
29. Medsker L.R., Jain L. (Eds.). *Recurrent Neural Networks. Design and Applications*. Boca Raton: CRC Press; 2001. 416 p.
30. Li J., Lavrukhin V., Ginsburg B., Leary R., Kuchaiev O., Cohen J.M., Nguyen H., Gadde R.T. Jasper: An End-to-End Convolutional Neural Acoustic Model. *Interspeech 2019*. 2019. https://doi.org/10.21437/interspeech.2019-1819
31. Dawalatabad N., Ravanelli M., Grondin F., Thienpondt J., Desplanques B., Na H. *ECAPA-TDNN Embeddings for Speaker Diarization.* arXiv preprint arXiv:2104.01466. 2021. https://doi.org/10.48550/arXiv.2104.01466
32. Ravanelli M., Bengio Y. Speaker recognition from raw waveform with SincNet. In: *2018 IEEE Spoken Language Technology Workshop* (*SLT*). IEEE; 2018. P. 1021–1028. https://doi.org/10.1109/SLT.2018.8639585
33. Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S. X-vectors: Robust DNN embeddings for speaker recognition. In: *2018 IEEE international Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2018. P. 5329–5333. http://doi.org/10.1109/ICASSP.2018.8461375
34. Deng J., Guo J., Xue N., Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2019. P. 4690–4699. https://doi.org/10.1109/CVPR.2019.00482
35. Gulati A., Qin J., Chiu C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. Conformer: Convolution-augmented trans-former for speech recognition. *Proc. Interspeech 2020*. 2020. P. 5036–5040. https://doi.org/10.21437/interspeech.2020-3015
36. Glorot X., Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*. 2010. P. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html

37. Gu J., Wang Z., Kuen J., Ma L., Shahroudy A., Shuai B., Liu T., Wang X., Wang G., Cai J., Chen T. Recent advances in convolutional neural networks. *Pattern Recognition*. 2018;77:354–377. https://doi.org/10.1016/j.patcog.2017.10.013
38. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017;30:5999–6009. https://doi.org/10.48550/arXiv.1706.03762
39. Ren Y., Ruan Y., Tan X., Qin T., Zhao S., Zhao Z., Liu T.Y. FastSpeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*. 2019;32. https://doi.org/10.48550/arXiv.1905.09263
40. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: *Proceedings of the 23rd International Conference on Machine Learning*. 2006. P. 369–376. https://doi.org/10.1145/1143844.1143891
41. Ghahremani P., BabaAli B., Povey D., Riedhammer K., Trmal J., Khudanpur S. A pitch extraction algorithm tuned for automatic speech recognition. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2014. P. 2494–2498. http://doi.org/10.1109/ICASSP.2014.6854049
42. Gerhard D. *Pitch Extraction and Fundamental Frequency: History and Current Techniques*. Masters Thesis. Regina, SK, Canada: Department of Computer Science, University of Regina; 2003. 23 p.
43. Łańcucki A. Fastpitch: Parallel text-to-speech with pitch prediction. In: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2021. P. 6588–6592. https://doi.org/10.1109/ICASSP39728.2021.9413889
44. Kong J., Kim J., Bae J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*. 2020;33:17022–17033. http://doi.org/10.48550/arXiv.2010.05646
45. Kominek J., Black A.W. The CMU Arctic speech databases. In: *Fifth ISCA Workshop on Speech Synthesis*. 2004. P. 223–224.
46. Zhao G., Sonsaat S., Silpachai A., Lucic I., Chukharev-Hudilainen E., Levis J., Gutierrez-Osuna R. L2-ARCTIC: A Non-native English Speech Corpus. *Interspeech 2018*. 2018. P. 2783–2787. http://doi.org/10.21437/Interspeech.2018-1110
47. Weinberger S.H., Kunath S.A. The Speech Accent Archive: towards a typology of English accents. In: *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Brill; 2011. P. 265–281. https://doi.org/10.1163/9789401206884_014
48. Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., Henretty M., Morais R., Saunders L., Tyers F., Weber G. Common Voice: A Massively-Multilingual Speech Corpus. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020. P. 4218–4222. https://doi.org/10.48550/arXiv.1912.06670
49. Nagrani A., Chung J.S., Zisserman A. Voxceleb: a large-scale speaker identification dataset. *Interspeech 2017*. 2017. http://doi.org/10.21437/Interspeech.2017-950
50. Chung J., Nagrani A., Zisserman A. VoxCeleb2: Deep speaker recognition. *Interspeech 2018*. 2018. http://doi.org/10.21437/Interspeech.2018-1929
51. Panayotov V., Chen G., Povey D., Khudanpur S. Librispeech: an ASR corpus based on public domain audio books. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2015. P. 5206–5210. http://doi.org/10.1109/ICASSP.2015.7178964
52. Veaux C., Yamagishi J., MacDonald K. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. *University of Edinburgh. The Center for Speech Technology Research* (*CSTR*). 2017. https://doi.org/10.7488/ds/2645
53. Chen G., Chai S., Wang G., Du J., Zhang W., Weng C., Su D., Povey D., Trmal J., Zhang J., Jin M., Khudanpur S., Watanabe S., Zhao S., Zou W., Li X., Yao X., Wang Y., Wang Y., You Z., Yan Z. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In: *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021*. International Speech Communication Association; 2021. P. 4376–4380. https://doi.org/10.21437/Interspeech.2021-1965
54. Bakhturina E., Zhang Y., Ginsburg B. Shallow Fusion of Weighted Finite-State Transducer and Language Model for Text Normalization. *Proc. Interspeech 2022*. 2022. http://doi.org/10.48550/arXiv.2203.15917
55. Kudo T., Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2018. P. 66–71. https://doi.org/10.48550/arXiv.1808.06226
56. Koizumi Y., Zen H., Karita S., Ding Y., Yatabe K., Morioka N., Bacchiani M., Zhang Y., Han W., Bapna A. Libritts-r: A Restored Multi-Speaker Text-to-Speech Corpus. *arXiv preprint arXiv:2305.18802*. 2023. https://doi.org/10.48550/arXiv.2305.18802
57. Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., Desmaison A., Kopf A., Yang E., DeVito Z., Raison M., Tejani A., Chilamkurthy S., Steiner B., Fang L., Bai J., Chintala S. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*. 2019;32:8024–8035.
58. Kuchaiev O., Li J., Nguyen H., Hrinchuk O., Leary R., Ginsburg B., Kriman S., Beliaev S., Lavrukhin V., Cook J., Castonguay P., Popova M., Huang J., Cohen J. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*. 2019. https://doi.org/10.48550/arXiv.1909.09577
59. Bredin H., Yin R., Coria J.M., Gelly G., Korshunov P., Lavechin M., Fustes D., Titeux H., Bouaziz W., Gill M.P. Pyannote. Audio: neural building blocks for speaker diarization. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2020. P. 7124–7128. https://doi.org/10.1109/ICASSP40776.2020.9052974

60. Olatunji T., Afonja T., Yadavalli A., Emezue C.C., Singh S., Dossou B.F., Osuchukwu J., Osei S., Tonja A.L., Etori N., Mbataku C. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics*. 2023;11:1669–1685. https://doi.org/10.1162/tacl_a_00627
61. Majumdar S., Balam J., Hrinchuk O., Lavrukhin V., Noroozi V., Ginsburg B. Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition. *arXiv preprint arXiv:2104.01721*. 2021. http://doi.org/10.48550/arXiv.2104.01721
62. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR 202. 2023. P. 28492–28518. http://doi.org/10.48550/arXiv.2212.04356

Об авторах

Нечаев Владимир Алексеевич, преподаватель-исследователь, ФГБОУ ВО «Ивановский государственный энергетический университет имени В.И. Ленина» (153003, Россия, Иваново, ул. Рабфаковская, д. 34). E-mail: nechaev@gapps.ispu.ru. SPIN-код РИНЦ 7002-3878, https://orcid.org/0009-0007-1449-3968

Косяков Сергей Витальевич, д.т.н., профессор, заведующий кафедрой программного обеспечения компьютерных систем, ФГБОУ ВО «Ивановский государственный энергетический университет имени В.И. Ленина» (153003, Россия, Иваново, ул. Рабфаковская, д. 34). E-mail: ksv@ispu.ru. Scopus Author ID 6507182528, ResearcherID H-5686-2018, SPIN-код РИНЦ 1371-9929, https://orcid.org/0000-0003-0231-0750

About the Authors

Vladimir A. Nechaev, Teacher-Researcher, Ivanovo State Power Engineering University (34, Rabfakovskaya ul., Ivanovo, 153003 Russia). E-mail: nechaev@gapps.ispu.ru. RSCI SPIN-code 7002-3878, https://orcid.org/0009-0007-1449-3968

Sergey V. Kosyakov, Dr. Sci. (Eng.), Professor, Head of the Department of Computer Systems Software, Ivanovo State Power Engineering University (34, Rabfakovskaya ul., Ivanovo, 153003 Russia). E-mail: ksv@ispu.ru. Scopus Author ID 6507182528, ResearcherID H-5686-2018, RSCI SPIN-code 1371-9929, https://orcid.org/0000-0003-0231-0750