УДК 004.934.2:004.032.26

НЕАВТОРЕГРЕССИОННАЯ МОДУЛЬНАЯ СИСТЕМА СИНТЕЗА РЕЧИ С ЯВНЫМ МОДЕЛИРОВАНИЕМ ДЛИТЕЛЬНОСТИ НА БАЗЕ CTC-ВЫРАВНИВАНИЯ

С.С. Бекарев
Студент факультета компьютерных систем и сетей БГУИР
bekarevstanislav@gmail.com

М.А. Калугина
Доцент кафедры информатики, кандидат физико-математических наук, доцент
marina_kalugina@list.ru

С.С. Бекарев
В 2022 году окончил гимназию №1 г. Горки. Интересы направлены на глубокое обучение, обучение моделей синтеза и распознавания речи.

М.А. Калугина
Окончила Белорусский государственный университет. Область научных интересов связана с исследованием проблем метрической теории диофантовых приближений зависимых величин и приложений математических методов к нейросетевому анализу

Аннотация. В статье рассматриваются методы явного выравнивания текстовых и акустических последовательностей для стабильного синтеза речи на материале русского языка. Проведено построение выравнивания текста относительно акустических признаков методом на основе CTC-loss и алгоритма Витерби. Для акустического моделирования использована неавторегрессионная архитектура с дилатационными свертками и энкодер-декодерной структурой. Оценка качества выполнена по группам показателей: точность предсказания длительности фонем (RMSE), качество реконструкции мел-спектрограмм (MSE) и субъективная разборчивость синтезированной речи (алгоритм Гриффина-Лима). Установлено, что предлагаемый трехэтапный пайплайн (обучение выравнивания, предсказание длительности графем, затем генерация мел-спектрограммы) обеспечивает устойчивое обучение и гарантированную монотонность генерации на графемном уровне с явной маркировкой ударения. Показана возможность обхода без фонемного преобразования при использовании расширенного словаря графем с ударными/безударными вариантами.

Ключевые слова: text-to-speech, CTC-loss, алгоритм Витерби, ASR, мел-спектрограмма, графемы, акустическая модель, алгоритм Гриффина-Лима, корпус RUSLAN.

Введение. Современные системы нейросетевого синтеза речи переходят от авторегрессионных архитектур (Tacotron [1], WaveNet [2]) к неавторегрессионным моделям, обеспечивающим ускорение инференса и повышение стабильности генерации. Авторегрессионные подходы, при всем высоком качестве синтеза, наследуют фундаментальные ограничения: накопление ошибок при экспозиционном смещении в условиях teacher forcing [3], а также непредсказуемое время генерации из-за последовательной генерации. Целью исследования является обучение модульной неавторегрессионной системы синтеза речи, использующей явное выравнивание текста и аудио для предсказания длительности фонем с последующей параллельной генерацией акустических признаков. Ключевой особенностью работы является реализация двухэтапного пайплайна на базе CTC-выравнивания: сначала извлекается строгий монотонный alignment между графемами и кадрами мел-спектрограммы с помощью алгоритма Витерби [4], затем на основе полученных

меток обучается предиктор длительности (DurationPredictor), обеспечивающий расширение последовательности эмбеддингов до акустического разрешения без использования attention-механизма. В работе представлена реализация полного цикла обучения системы на корпусе русской речи RUSLAN [6] (30 часов, один спикер): от подготовки графемной транскрипции с явным кодированием ударения (токены вида «+а», «+е») до генерации мел-спектрограмм [7] (80 мел-фильтров) с использованием акустической модели на основе Unet-like архитектуры [8]. Оценка качества проводится по метрикам точности предсказания длительности (RMSE), реконструкции спектральных признаков (MSE) и субъективной разборчивости синтезированной речи, полученной алгоритмом Гриффина-Лима [9].

Описание алгоритма обучения. В качестве источника акустических данных используется открытый корпус русской речи RUSLAN с текстовыми транскрипциями. Такой подход обеспечивает фиксацию акустических характеристик конкретного спикера без вариативности, свойственной многоголосым датасетам. На этапе предобработки аудиозаписи преобразуются в мел-спектрограммы с параметрами: 80 мел-фильтров, окно Ханна длительностью 64 мс, сдвиг 16 мс, логарифмическое шкалирование (log1p) для стабилизации динамического диапазона. Текстовые транскрипции токенизируются на уровне графем с явным кодированием ударения: каждая гласная представлена двумя токенами (ударная с префиксом «+» и безударная).

Для получения целевых меток длительности обучается вспомогательная сверточная модель автоматического распознавания речи с использованием CTC-loss [10].

После достижения CER 9% к выходам сети применяется алгоритм Витерби для извлечения оптимального монотонного alignment, устанавливающего соответствие между каждым токеном текста и фреймами мел-спектрограммы. Полученные alignment'ы служат эталоном для обучения DurationPredictor. Каждый пример обучающей выборки включает последовательность токенов графем с маркировкой ударения и целевой вывод: вектор длительностей (количество фреймов для каждого токена) и соответствующий фрагмент мел-спектрограммы.

Для формирования обучающей пары для AcousticModel текстовая последовательность расширяется путем повторения эмбеддингов токенов, обеспечивая временное соответствие между входом и целевой спектрограммой.

Процесс обучения выполняется на GPU в три последовательных этапа: (1) обучение ASR-модели с CTC-loss до сходимости; (2) обучение DurationPredictor на извлеченных алгоритмом Витерби alignment'ах с функцией потерь MSE; (3) обучение акустической модели на сформированных расширенных эмбеддингах с целевыми мел-спектрограммами (MSELoss). Контроль качества осуществляется на валидационной выборке по метрикам точности предсказания длительности (RMSE в кадрах), качества реконструкции спектральных признаков (MSE), а также субъективной разборчивости речи, синтезированной из предсказанных спектрограмм с помощью алгоритма Гриффина-Лима.

Поскольку записи в корпусе RUSLAN имеют неравную длительность (от коротких аудио до длинных предложений из десятка слов), формирование батчей требует выравнивания последовательностей до единой длины внутри батча. Для предотвращения влияния искусственно добавленных элементов (padding) на градиенты модели применяется механизм маскирования (masking).

При обучении ASR-модели длины входных и целевых последовательностей передаются в CTC-loss, который автоматически игнорирует вклад padding-токенов при расчете log-likelihood. Для DurationPredictor и AcousticModel используется бинарная маска, идентифицирующая позиции реальных данных внутри выровненных тензоров, при вычислении MSELoss маска применяется поэлементно, обнуляя вклад фиктивных позиций в функцию потерь и обеспечивая корректное усреднение только по актуальным временным шагам.

Результаты обучения системы. На первом этапе реализована ASR-модель для распознавания графем (включая маркеры ударения).

Применение алгоритма Витерби к выходам сети позволило извлечь строгий монотонный путь, установивший однозначное соответствие между токенами текста и фреймами мел-спектрограммы. Пример alignment'а текста можно увидеть на рисунке 1.

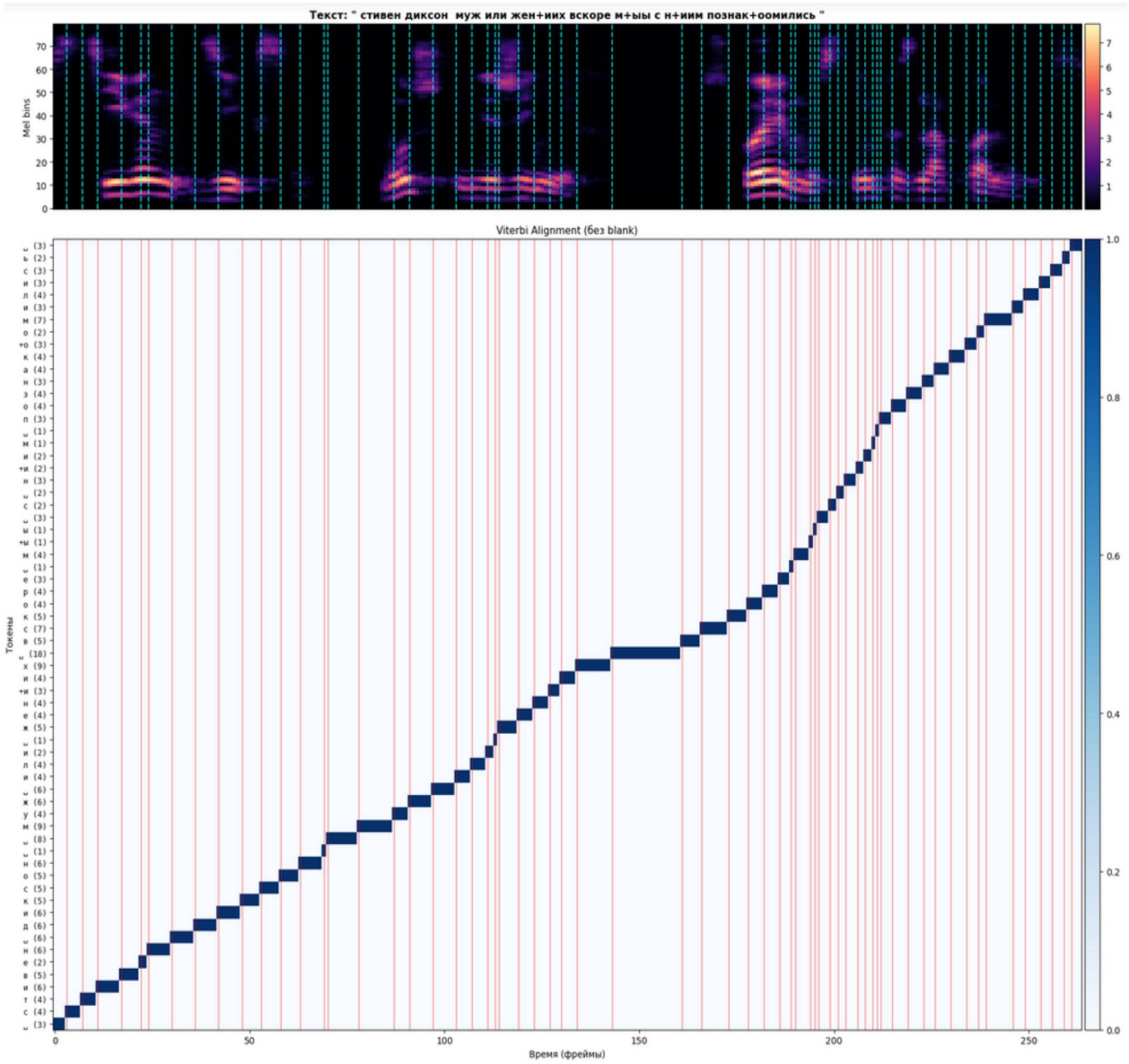

Рисунок 1. Alignment текста, полученный с помощью алгоритма Витерби

На основе извлеченных alignment'ов был обучен DurationPredictor. Кривые обучения демонстрируют монотонное снижение MSELoss: с начального значения 80 до 44 на 12-й эпохе, то есть RMSE 6.6 на валидационной выборке (рисунок 2).

Достаточно высокое значение RMSE можно объяснить тем, что нейронная сеть недооценивает длину пауз между произношением слов текста, порой пауза может доходить до 30 фреймов мел-спектрограммы, что негативно сказывается на общей метрике.

Для предотвращения переобучения, обучение было заранее остановлено.

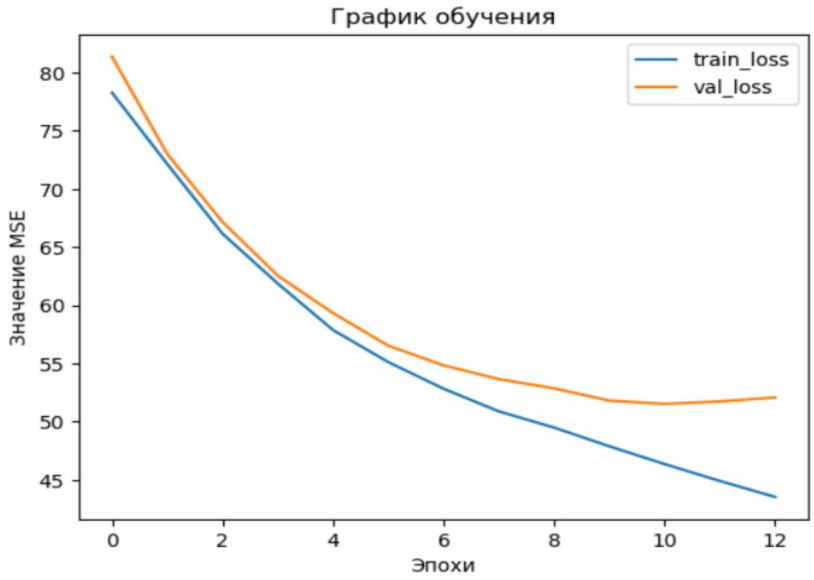

Рисунок 2. Динамика MSE предсказания длительности для DurationPredictor

Финальный этап – обучение акустической модели на архитектуре с дилатационными свертками (dilated convolutions) и стратегией downsampling/upsampling (Unet-like архитектура). Модель принимает на вход расширенные эмбеддинги графем (согласно предсказанным длительностям) и генерирует логарифмированные мел-спектрограммы (80 мел-фильтров). Обучение проводилось с MSELoss, значение которого снизилось до 0.5 на 18-й эпохе.

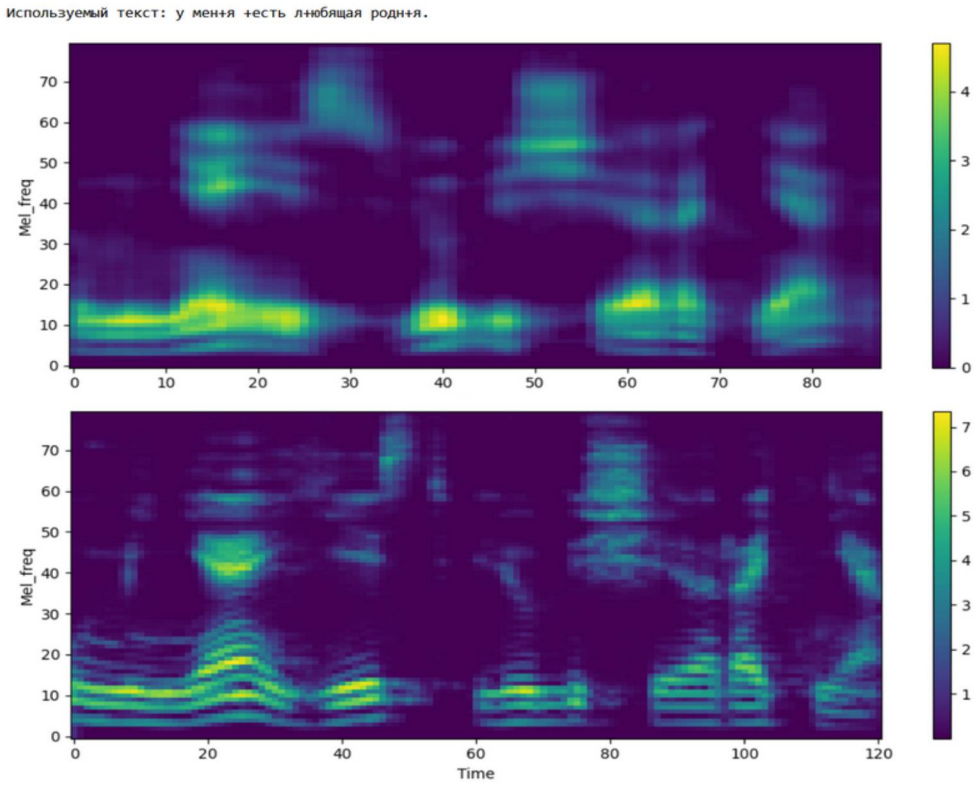

Рисунок 3. Сравнение спектрограмм: синтез DurationPredictor и AcousticModel (сверху), оригинальная запись (снизу)

Использование фиксированной, детерминированной длины последовательности (задаваемой алгоритмом Витерби) обеспечило устойчивую сходимость без колебаний функции потерь, что соответствует качеству, достаточному для разборчивости речи при использовании вокодера Griffin-Lim (рисунок 3).

Из сравнения видно, что предсказания DurationPredictor получаются немного смещенными, а предсказания AcousticModel не обладают высокой резкостью, от чего мел-спектрограмма получается сглаженной, но общая структура звучания текста сохраняется, что позволяет разобрать речь на аудио, полученном после преобразования Griffin-Lim.

Таким образом, реализованный трехэтапный пайплайн обеспечил устойчивое обучение всех компонентов системы.

Достигнутые показатели подтверждают работоспособность подхода для синтеза речи на основе графемной транскрипции с маркировкой ударения без привлечения фонематора.

Заключение. В ходе исследования разработана и экспериментально проверена методика построения модульной системы синтеза речи, использующей явное выравнивание текста и аудио для обеспечения стабильности обучения non-autoregressive генерации акустических признаков.

Предложенный подход включает обучение ASR-модели с CTC-loss, извлечение монотонного alignment алгоритмом Витерби, предсказание длительности фонем (DurationPredictor) и параллельную генерацию мел-спектрограмм акустической моделью на базе дилатационных сверток.

Проведенный анализ архитектурных решений подтвердил, что выбор метода выравнивания определяющим образом влияет на сходимость обучения и качество синтеза речи на материале русского языка:

1 Метод явного CTC-выравнивания с алгоритмом Витерби обеспечил гарантированную монотонность отображения графем на акустическую ось и устойчивую сходимость всех компонентов системы.

Трехэтапная схема (выравнивание → предсказание длительности → генерация спектрограмм) продемонстрировала стабильное снижение функций потерь без признаков переобучения: RMSE предсказания длительности достиг 6.6 кадра, MSE реконструкции мел-спектрограмм – 0.5.

При этом достигнута разборчивость текста при синтезе через алгоритм Гриффина-Лима, что подтверждает работоспособность подхода.

2 Использование графемной записи с явной маркировкой ударения (вместо фонем) позволило упростить пайплайн предобработки текста, однако внесло зависимость от точности ASR-модели (CER 9%).

Ошибки распознавания транслируются в локальные искажения длительности, что ограничивает качество синтеза для сложных фонетических конструкций.

Таким образом, проведённое исследование показало, что предложенный модульный подход с CTC-выравниванием обеспечивает детерминированность генерации, фиксированное время инференса моделей и устойчивость к переобучению при ограниченном объеме данных (30 часов).

При этом сохраняется возможность дальнейшего улучшения качества путем замены отдельных компонентов (например, дообучение ASR для снижения CER) без перестроения всей системы.

Список литературы

[1] NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS [Электронный ресурс]. – Электронные данные. – Режим доступа: https://arxiv.org/pdf/1712.05884 – Дата доступа: 21.03.2026

[2] WAVENET: A GENERATIVE MODEL FOR RAW AUDIO [Электронный ресурс]. – Электронные данные. – Режим доступа: https://arxiv.org/pdf/1609.03499 – Дата доступа: 21.03.2026

[3] RNNs: Teacher Forcing [Электронный ресурс]. – Электронные данные. – Режим доступа: https://cedar.buffalo.edu/~srihari/CSE676/10.2.1%20TeacherForcing.pdf – Дата доступа: 21.03.2026

[4] Распознавание речи: Классический подход [Электронный ресурс]. – Электронные данные. – Режим доступа: http://www.machinelearning.ru/wiki/images/c/c3/Digital_Signal_ Processing%2C_lecture_6.pdf – Дата доступа: 21.03.2026

[5] Что такое ASR [Электронный ресурс]. – Электронные данные. – Режим доступа: https://microsin.net/programming/arm/what-is-asr.html – Дата доступа: 21.03.2026

[6] Hugging Face [Электронный ресурс]. – Электронные данные. – Режим доступа: https://huggingface.co/datasets/Gzaborey/ruslan-dataset – Дата доступа: 18.03.2026

[7] Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between [Электронный ресурс]. – Электронные данные. – Режим доступа: https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html – Дата доступа: 18.03.2026

[8] U-Net Architecture Explained [Электронный ресурс]. – Электронные данные. – Режим доступа: https://www.geeksforgeeks.org/machine-learning/u-net-architecture-explained/ – Дата доступа: 18.03.2026

[9] The Griffin-Lim algorithm: Signal estimation from modified short-time Fourier transform [Электронный ресурс] – Электронные данные – Режим доступа: https://speechprocessingbook.aalto.fi/Modelling/griffinlim.html – Дата доступа: 18.03.2026

[10] Распознавание речи: Современные подходы [Электронный ресурс]. – Электронные данные. – Режим доступа: http://www.machinelearning.ru/wiki/images/c/c6/Digital_Signal_ Processing%2C_lecture_7.pdf – Дата доступа: 21.03.2026.

Авторский вклад

Бекарев Станислав Сергеевич – выбор задачи исследования, формулировка концепции трехэтапного пайплайна обучения (CTC-выравнивание → предсказание длительности → генерация спектрограмм); разработка и реализация всех компонентов системы (ASR-модель, DurationPredictor, AcousticModel); создание инструментов визуализации alignment'ов; проведение экспериментов на корпусе RUSLAN и анализ метрик.

Калугина Марина Алексеевна – научное руководство исследованием; постановка проблемы; консультации по математическим основам CTC-loss и алгоритмов выравнивания.

NON-AUTOREGRESSIVE MODULAR SPEECH SYNTHESIS SYSTEM WITH EXPLICIT DURATION MODELING BASED ON CTC EQUALIZATION

S.S. Bekarev
Student of BSUIR

M.A. Kalugina
Associate Professor of Informatics Department of the BSUIR

Abstract. This article examines methods for explicitly aligning text and acoustic sequences for stable speech synthesis using Russian language material. Emphasis is placed on overcoming the instability of attention mechanisms in classical seq2seq models by decomposing the problem into alignment, duration prediction, and spectrogram generation modules. A text alignment with respect to acoustic features is constructed using a method based on CTC-loss and the Viterbi algorithm. A non-autoregressive architecture with dilated convolutions and an encoder-decoder structure is used for acoustic modeling. Quality assessment is performed using a group of metrics: phoneme duration prediction accuracy (RMSE), mel spectrogram reconstruction quality (MSE), and subjective intelligibility of synthesized speech (Griffin-Lim algorithm). It was established that the proposed two-stage pipeline (grapheme duration prediction followed by mel-spectrogram generation) ensures robust learning and guaranteed monotonicity of generation at the grapheme level with explicit stress marking. The feasibility of traversal without phoneme transformation is demonstrated using an extended grapheme dictionary with stressed/unstressed variants.

Keywords: text-to-speech, CTC-loss, Viterbi algorithm, non-autoregressive TTS, ASR, mel-spectrogram, graphemes, dilated convolutions, acoustic model, Griffin-Lim algorithm, RUSLAN corpus.