АРХИТЕКТУРА NON-AUTOREGRESSIVE СИНТЕЗА РЕЧИ С CTC-ВЫРАВНИВАНИЕМ НА ОСНОВЕ ДВУХЭТАПНОГО ПАЙПЛАЙНА

Бекарев С.С., студент

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Калугина М.А. – канд. физ.-мат. наук, доцент

Аннотация. В работе представлена модульная архитектура синтеза речи, комбинирующая методы автоматического распознавания речи для получения точного выравнивания текста и аудио с non-autoregressive генерацией акустических признаков. Для преодоления проблем нестабильности attention-механизмов, характерных для seq2seq моделей семейства Tacotron, разработан двухэтапный пайплайн: на первом этапе обучается CTC-based модель распознавания речи, после чего алгоритм Витерби применяется для извлечения строгого монотонного alignment между символами текста русского языка и кадрами мел-спектрограммы. Полученные временные метки используются для обучения отдельного DurationPredictor (предиктор длительности), предсказывающего количество фреймов для каждого входного символа. Акустическая модель на основе расширенных эмбеддингов генерирует логарифмированные мел-спектрограммы (80 мел-фильтров).

Современные системы синтеза речи демонстрируют стремительное развитие в направлении создания естественно звучащей речи, однако переход от лабораторных прототипов к промышленным внедрениям сопряжен с фундаментальными проблемами стабильности генерации. Традиционные end-to-end архитектуры, опирающиеся на механизм внимания (attention) для неявного выравнивания текста и акустических признаков, несмотря на высокое качество результата, характеризуются непредсказуемостью поведения при обучении: нестабильность монотонности alignment, пропуск символов и характерные артефакты воспроизведения затрудняют их применение в задачах, критичных к надежности. Особую сложность представляет адаптация таких систем к морфологически богатым языкам, таким как русский, где фонетическая реализация графем существенно зависит от позиции ударения.

Актуальность исследования обусловлена необходимостью преодоления принципиального ограничения attention-based моделей, связанного с отсутствием гарантий монотонности отображения последовательности символов во временную ось. Существующие подходы требуют сложных эвристик (guided attention loss, windowing) и тонкой настройки гиперпараметров, при этом процесс обучения остается восприимчивым к сбоям на сложных фонетических конструкциях. Для русского языка, где длительность звуков сильно варьируется в зависимости от ударения, отсутствие явного контроля над временной разверткой текста приводит к неестественной ритмике и дефектам произношения, что ограничивает применимость технологий в синтезаторах речи профессионального уровня.

В связи с этим перспективным направлением является отказ от неявного alignment в пользу явного моделирования длительности фонем на основе предварительного выравнивания текста и аудио. Применение методов автоматического распознавания речи (ASR) с CTC-loss [1] и алгоритмов динамического программирования (Viterbi) [2] для извлечения точных временных меток позволяет декомпозировать задачу на управляемые этапы: сначала определение оптимального пути alignment, затем предсказание длительности на основе полученных данных, и наконец генерация акустических признаков. Подобная модульная архитектура обеспечивает стабильность обучения и контролируемую генерацию, сохраняя при этом высокое качество синтеза.

В данной работе приведены результаты разработки non-autoregressive пайплайна синтеза речи, идея которого заключается в использовании CTC-выравнивания для обучения DurationPredictor и последующей генерации мел-спектрограмм акустической моделью без привлечения механизма attention. В основе подхода лежит двухэтапная схема: обучение ASR-модели (CNN с Conv1D) на корпусе RUSLAN [3] с последующим применением алгоритма Витерби для получения строгого монотонного alignment (CER 9%), на базе которого обучается предиктор длительности, и акустическая модель на архитектуре энкодер-декодер с дилатационными свертками, работающая с графемной записью текста с явной маркировкой ударений.

На первом этапе реализована система автоматического распознавания речи (ASR) на архитектуре сверточной нейронной сети с одномерными свертками (Conv1D). Модель обучается с использованием CTC-loss (Connectionist Temporal Classification), который обеспечивает дифференцируемое выравнивание последовательностей различной длины без явного указания временных границ. Для получения строгого монотонного alignment применяется алгоритм Витерби, находящий оптимальный путь через матрицу posterior probabilities методом динамического программирования. Достигнутая точность распознавания составляет CER 9%, что является достаточным для извлечения надежных временных меток.

Полученный alignment используются для обучения DurationPredictor — полносвязной нейронной сети, предсказывающей количество фреймов мел-спектрограммы для каждого входного символа. На вход подаются эмбеддинги графем (включая варианты с ударением) без дополнительных позиционных

кодировок или контекстуальных признаков. Обучение проводится с функцией потерь MSE, минимизирующей среднеквадратичное отклонение предсказанной длительности от эталонной, полученной из Viterbi-alignment.

Акустическая модель построена по энкодер-декодерной схеме с использованием дилатационных сверток (dilated 1d convolutions). Архитектура включает блоки последовательного downsampling (для сжатия временного разрешения и увеличения рецептивного поля) и upsampling (восстановление исходной длины последовательности). Данная структура обеспечивает эффективное моделирование долгосрочных зависимостей в акустических данных при фиксированной длине входной последовательности, определяемой выходом DurationPredictor. Модель генерирует логарифмированные мел-спектрограммы, минимизируя MSE относительно оригинальных акустических признаков (оригинальной мел-спектрограммы).

Оценка качества производилась по метрикам: MSE предсказания длительности (кадры), MSE реконструкции мел-спектрограммы, а также субъективно — по разборчивости и естественности речи, синтезированной алгоритмом Гриффина-Лима (Griffin-Lim) [4]. Ниже представлена таблица с результатами оценки.

Таблица 1 – Результаты оценки модульного подхода генерации акустических признаков

Критерий оценки	Результат модели
CER (ASRModel)	9%
RMSE (DurationPredictor)	7.14
MSE (AcousticModel)	0.26
Разборчивость речи	выше среднего
Естественность речи	средняя

Вследствие проведения экспериментов также установлено, что предлагаемый модульный подход обеспечивает устойчивое обучение в условиях отсутствия предварительной фонемной транскрипции. Базовая Tacotron-модель продемонстрировала нестабильность attention-механизма даже при использовании teacher forcing и guided attention loss (несходимость attention-механизма). Альтернативный метод с линейным растягиванием текста в эмбединговом пространстве символов и soft-DTW loss, хотя и обеспечил сходимость, показал существенно более высокие значения MSE при генерации акустических признаков, что объясняется неспособностью линейной интерполяции адекватно моделировать нелинейные соответствия между текстовой и акустической последовательностями.

Предлагаемая архитектура достигла приемлемого для vocoder-free синтеза (синтез с помощью алгоритма Griffin-Lim), обеспечивая разборчивость текста при характерном «роботизированном» тембре, связанном с отсутствием фазовой информации. Отсутствие артефактов перескока букв подтверждает корректность использования явного CTC-выравнивания и Viterbi-алгоритма.

Проведенное исследование подтвердило, что декомпозиция задачи синтеза речи на этапы выравнивания, предсказания длительности и генерации акустических признаков является эффективной альтернативой end-to-end подходам. Использование графемной записи с маркировкой ударения позволяет обходиться без фонематора.

Список использованных источников:

1. Breaking down the CTC Loss [Электронный ресурс]. – Электронные данные. – Режим доступа: https://ogunlao.github.io/blog/2020/07/17/breaking-down-ctc-loss.html – Дата доступа: 18.03.2026

2. Распознавание речи [Электронный ресурс]. – Электронные данные. – Режим доступа: http://www.machinelearning.ru/wiki/images/c/c3/Digital_Signal_Processing%2C_lecture_6.pdf – Дата доступа: 18.03.2026

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

4. PyTorch Documentation [Электронный ресурс]. – Электронные данные. – Режим доступа: https://docs.pytorch.org/audio/2.8/generated/torchaudio.transforms.GriffinLim.html – Дата доступа: 18.03.2026