МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«Тольяттинский государственный университет»

Кафедра «Прикладная математика и информатика»
(наименование)

09.04.03 Прикладная информатика
(код и наименование направления подготовки)

Управление корпоративными информационными процессами
(направленность (профиль))

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)

на тему Алгоритмы синтеза речи с применением глубокого обучения для создания речевых моделей, имитирующих голоса конкретных людей

Обучающийся И. А. Варичев
(Инициалы Фамилия) (личная подпись)

Научный руководитель канд. пед. наук, доцент О. В. Оськина
(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Тольятти 2025

Содержание

Введение 3
1 Теоретические основы и современные подходы к синтезу речи 5
1.1 Современное состояние научной проблемы 5
1.2 Обоснование актуальности исследования 9
1.3 Обзор технологий TTS синтеза речи 11
1.4 Обзор технологий VC синтеза речи 14
1.5 Обоснование выбора объекта и предмета исследования 18
2 Сравнительный анализ алгоритмов синтеза речи 24
2.1 Критерии и метрики сравнения алгоритмов 24
2.2 Сравнительный анализ TTS моделей. 27
2.3 Сравниетельный анализ VC моделей. 35
2.4 Оценка влияния интеграций на качество синтеза 47
3 Проектирование конвейера синтеза речи 50
3.1 Проектирование диаграммы гибридной модели 50
3.2 Описание модулей и их задач 54
4 Анализ результатов и оценка эффективности конвейера 85
4.1 Анализ результатов исследования 86
4.2 Оценка вычислительной сложности и латентности 88
4.3 Анализ альтернативных решений 96
4.4 Сравнительный анализ эффективности предложенной модели 102
4.5 Практические и методологические рекомендации 107
Заключение 111
Список используемой литературы и используемых источников 113

Введение

Синтез речи применяется повсеместно, однако существующие решения сталкиваются с ограничениями: высокий уровень фонетических ошибок, потеря индивидуальных характеристик голоса, неэффективность обучения и инференса, а также отсутствие единого промежуточного представления для согласованного взаимодействия модулей TTS и VC.

Разрозненные этапы Text-to-Speech и Voice Conversion приводят к утрате просодических нюансов и снижению естественности синтезированной речи, особенно в условиях ограниченных обучающих данных.

Цель исследования, разработать концепцию интегрированной модульной модели синтеза речи, объединяющей алгоритмические подходы TTS и VC в единую систему.

Объект исследования: совокупность процессов синтеза речи.

Предмет исследования: анализ и оценка интеграционных стратегий для объединения алгоритмических подходов TTS и VC в единую модель синтеза речи.

Гипотеза исследования: интеграция TTS и VC с использованием self-supervised learning обеспечивает более эффективный синтез речи по сравнению с раздельным применением этих технологий.

Задачи исследования:

- Изучить современные алгоритмические методы TTS и VC, оценить их преимущества и ограничения.
- Систематизировать проблемы разрозненной обработки и выявить ключевые вопросы интеграции.
- Обосновать эффективность единого конвейерного подхода для повышения качества и адаптивности.
- Разработать концепцию модульной интеграции TTS и VC.

– Сформулировать методические рекомендации для дальнейших исследований и внедрения.

Основные этапы исследования и его апробация:

– Анализ литературы и формулировка критериев сравнительного анализа;
– Сравнительный анализ существующих TTS- и VC-моделей;
– Проектирование пятимодульного конвейера синтеза речи (TTS→VC)+SSL;
– Экспериментальная оценка производительности, качества и латентности;
– Выработка рекомендаций по интеграции и промышленному внедрению.

Научная новизна: предложена интеграция TTS и VC модулей через SSL.

Практическая значимость: определены оптимальные алгоритмы и модели, что позволяет улучшить показатели, снизить задержку и выявить узкие места.

Основные положения, выносимые на защиту:

– Сравнительный анализ существующих алгоритмов и моделей TTS и VC;
– Схема 5-модульного конвейера (TTS→VC)+SSL;
– Оценка производительности и качества.

Диссертация состоит из введения, четырёх глав, заключения и списка литературы:

– Глава 1 «Теоретические основы и современные подходы к синтезу речи»;
– Глава 2 «Сравнительный анализ алгоритмов TTS и VC»;
– Глава 3 «Проектирование гибридной модели синтеза речи»;
– Глава 4 «Анализ результатов и оценка эффективности предложенной модели».

1 Теоретические основы и современные подходы к синтезу речи

В первой главе раскрываются основные теоретические подходы к синтезу речи: от классических методов конкатенативного и параметрического синтеза до современных нейронных архитектур. Описываются принципы работы TTS и VC-систем, а также возможности и ограничения self-supervised обучения в задачах генерации и преобразования голоса.

1.1 Современное состояние научной проблемы

Технологии синтеза речи претерпели значительные изменения. Развитие синтеза речи опиралось на два основных подхода: конкатенативный синтез и статистический параметрический синтез.

Конкатенативный синтез речи основывается на использовании заранее записанных звуковых единиц для генерации речи, что позволяет получать достаточно естественное звучание. В этом подходе качество синтезированной речи определяется, прежде всего, консистентностью акустических характеристик на границах соединения отдельных единиц. Исследования показывают, что использование более крупных единиц, таких как слова или предложения, способствует получению более натурального звучания, однако ограничивает гибкость системы, снижая её способность адаптироваться к изменениям интонации или эмоциональной окраски. В то же время «применение более мелких единиц, таких как фонемы, обеспечивает большую гибкость, но может негативно сказаться на качестве синтеза речи» [29].

Статистический параметрический синтез речи основывается на использовании вероятностных моделей, для представления речевого сигнала через набор параметров. Этот подход обеспечивает большую гибкость генерации речи по сравнению с конкатенативным синтезом, поскольку позволяет адаптировать

параметры для достижения различных интонационных и стилевых эффектов. Несмотря на возможность детальной настройки данный подход отличался недостаточной естественностью звучания и ограниченной выразительностью [9].

Современные разработки в области синтеза речи основаны на нейронных архитектурах, реализующих технологии Text-to-Speech (TTS) и Voice Conversion (VC). Данные подходы, позволяют существенно повысить качество синтеза за счёт более точного моделирования временных зависимостей, передачи просодических характеристик и адаптации к новым условиям. В TTS используются современные алгоритмические подходы, основанные на глубоких нейронных сетях. Тем не менее, современные TTS-системы продолжают сталкиваться с рядом ограничений, которые существенно влияют на их практическую применимость. «Качество синтезируемой речи недостаточно естественно, адаптация систем к различным акцентам и стилям речи требует значительных объемов высококачественных обучающих данных, что усложняет их использование в условиях ограниченных ресурсов, при этом высокая вычислительная сложность современных моделей может приводить к задержкам» [41].

В области преобразования голоса применяются современные алгоритмические подходы, основанные на нейронных автоэнкодерах и генеративных состязательных сетях. Эти методы позволяют преобразовывать голосовые характеристики исходного говорящего в характеристики целевого, при этом сохраняется смысловое содержание речи.

В то же время, несмотря на значительный прогресс, технологии VC остаются подверженными ряду проблем. Во-первых, синтезированная речь зачастую не достигает уровня естественности исходного голоса из-за неточностей в передаче просодических и интонационных характеристик. Во-вторых, модели могут испытывать сложности с адаптацией к различным стилям и акцентам, что ограничивает их универсальность, особенно в условиях недостатка обучающих

данных. Кроме того, существует риск негативного переноса, когда адаптированные представления голоса не полностью соответствуют целевым характеристикам.

Интеграция современных алгоритмических подходов, включая использование DNN, LSTM, speaker-independent pre-training и обучение на непараллельных данных, значительно повышает естественность и адаптивность систем VC, хотя остаются проблемы точного воспроизведения просодических характеристик [39].

Исходя из анализа современных алгоритмических методов в TTS и VC, каждая технология обладает своими преимуществами, однако при разрозненном применении возникает риск потери важной информации между этапами обработки. Интеграция различных алгоритмических подходов в единую архитектуру является важным направлением, поскольку она обеспечивает согласованность и устойчивую передачу данных между этапами синтеза речи. Традиционно извлечение признаков, генерация акустических представлений и преобразование их в аудиосигнал выполнялись разрозненно, что приводило к потере информации и несогласованности результатов.

Современный метод self-supervised learning (SSL) играет ключевую роль в повышении качества синтеза речи. Self-supervised learning становится трансформирующей методологией, поскольку она позволяет обучать модели на неразмеченных данных, автоматически извлекая полезные представления, что повышает устойчивость системы и адаптивность к новым условиям, что приводит к улучшению качества, разнообразия и выразительности генерируемой речи. При этом, применение self-supervised learning в синтезе речи сопровождается рядом проблем. Использование этапа предобработки перед синтезом речи может приводить к накоплению ошибок, что приводит к несоответствию входных данных для последующего этапа. Кроме того, сам процесс предобработки зачастую оказывается потерьным, что может привести к утрате важных акустических признаков, необходимых для адекватного разделения речи и шума. Эффективность

представлений, получаемых методом self-supervised learning, в контексте улучшения синтеза речи также остаётся недостаточно изученной. Увеличение числа модулей предобработки дополнительно усложняет общую архитектуру системы, затрудняя её оптимизацию и интеграцию. «Объективные метрики, такие как STOI и PESQ, не всегда корректно отражают субъективное восприятие качества синтезируемой речи» [12], что подчёркивает необходимость дальнейшей оптимизации алгоритмов.

Интеграция self-supervised learning способствует непрерывной и корректной передаче данных между этапами синтеза, что позволяет создать универсальные и высококачественные системы синтеза речи. Такой подход обеспечивает, что результат работы одного модуля становится корректным входом для следующего этапа, а система может динамично обновляться без необходимости полного переобучения.

В итоге, современное состояние научной проблемы в области синтеза речи характеризуется тем, что, традиционные подходы, хоть и обеспечивали определённую естественность звучания, страдали от ограниченной гибкости и неспособности адекватно передавать тонкие просодические и эмоциональные характеристики. Современные нейронные методы, использующие модели Sequence-to-Sequence, неавторегрессионные архитектуры, механизмы внимания, а также современные техники обучения self-supervised learning, существенно улучшают качество синтеза, однако всё ещё остаются проблемы адаптивности, высоких вычислительных затрат и потери информации между этапами обработки. Кроме того, методы преобразования голоса (VC) сталкиваются с трудностями в точном воспроизведении просодии, адаптации к различным стилям и акцентам, а также с проблемами негативного переноса и недостаточной универсальностью. В совокупности, разрозненное применение этих методов подчеркивает необходимость разработки интегрированных архитектур, способных объединить

преимущества всех подходов для создания универсальных, устойчивых и высококачественных систем синтеза речи.

1.2 Обоснование актуальности исследования

В условиях стремительного развития технологий синтеза речи требования к качеству, адаптивности и естественности генерируемой речи постоянно растут. Несмотря на значительный прогресс, достигнутый переходом от традиционных методов, таких как конкатенативный и статистический параметрический синтез, к современным нейронным архитектурам, системы TTS и VC по-прежнему сталкиваются с рядом существенных ограничений.

Современные нейронные методы значительно повышают качество синтеза за счёт точного моделирования временных зависимостей и просодических характеристик, однако остаётся проблема адаптации к различным акцентам, стилям и языковым особенностям. Кроме того, высокая вычислительная сложность и разрозненное выполнение этапов обработки (извлечение признаков, генерация акустических представлений и преобразование их в аудиосигнал) приводят к потере важной информации, что негативно сказывается на итоговом качестве синтеза.

Методт self-supervised learning, открывают новые возможности для повышения устойчивости и адаптивности систем. Self-supervised learning автоматически извлекает полезные представления из неразмеченных данных, повышая устойчивость системы, однако вызывает вопросы каскадирования модулей и потерь информации на этапе предобработки, а также сложности в объективной оценке качества синтеза [25].

Современные методы синтеза речи показывают значительные успехи, однако их раздельное применение приводит к ряду проблем, ограничивающих качество и универсальность генерируемой речи. Несмотря на эффективность отдельных

алгоритмов, их взаимодействие в рамках единой системы остается нерешенной задачей. Основные проблемы современных подходов включают:

- Потеря информации между этапами обработки: в существующих системах TTS и VC каждый этап в большинстве случаев разрабатывается отдельно, что приводит к несовместимости представлений между модулями.
- Ограничения в адаптации к новым данным: многие современные алгоритмы требуют больших объемов размеченных данных для обучения. Это особенно актуально в задачах персонализированного синтеза речи, где необходимо адаптировать модель к конкретному голосу на основе небольшого количества примеров.
- Высокие вычислительные затраты: современные архитектуры требуют значительных ресурсов на этапе обучения и генерации, что делает сложным их применение в задачах реального времени.
- Ограниченная управляемость интонацией и просодией: точный контроль над интонацией, эмоциональной окраской и динамикой речи остается сложной задачей. Современные модели либо требуют явной разметки данных, либо предоставляют ограниченные механизмы управления характеристиками голоса.
- Недостаточная интеграция TTS и VC: системы синтеза речи и преобразования голоса в большинстве случаев рассматриваются отдельно, хотя их принципы работы во многом схожи. Отсутствие единого представления данных между этими задачами приводит к необходимости дублирования обработки на разных этапах, что увеличивает вычислительную сложность и снижает эффективность обработки.

Выявленные ограничения современных методов синтеза речи, применяемых разрозненно в системах TTS и VC, подчеркивают необходимость создания единой интегрированной модели, которая объединяет сильные стороны обоих направлений. В этом контексте предлагаются следующие стратегии и направления

оптимизации: Разработка модульной модели, основанной на общем промежуточном представлении, может обеспечить непрерывную и корректную передачу данных между этапами обработки.

Эти вызовы подчеркивают необходимость разработки интегрированных архитектур, способных объединить преимущества различных подходов в единую систему. Такой интегрированный подход обеспечивает непрерывную и корректную передачу данных между этапами синтеза, позволяя создать универсальные системы, которые не только улучшают естественность и адаптивность синтезируемой речи, но и снижают эксплуатационные затраты.

Актуальность исследования обусловлена необходимостью преодоления существующих ограничений традиционных и современных методов синтеза речи. Разработка интегрированного комбинированного подхода, основанного на объединении технологий TTS и VC с использованием передовых методов обучения, имеет существенное значение как для дальнейшего теоретического развития области, так и для практического применения голосовых технологий в условиях растущих требований к персонализации, мультиязычности и адаптивности.

1.3 Обзор технологий TTS синтеза речи

Современные системы синтеза речи опираются на разнообразные алгоритмические подходы, направленные на обеспечение качественного преобразования текста в речь и изменения голосовых характеристик.

В TTS используются архитектуры, которые преобразуют входной текст в акустические представления. Seq2Seq представляет собой архитектурный подход, предназначенный для преобразования последовательностей входных данных в последовательности выходных представлений, что является критически важным для синтеза речи. Seq2Seq обеспечивает гибкость обработки переменных последовательностей и возможность точного выравнивания текста с акустическими

характеристиками. Однако его применение сопряжено с высокими вычислительными затратами, а также ограничениями в передаче тонких нюансов просодии, особенно при обучении на ограниченных данных. Эти аспекты требуют дальнейшей оптимизации и являются важными направлениями для развития интегрированных систем синтеза речи [38].

End-to-End представляет собой интегрированные архитектуры, в которых все этапы преобразования входного текста в аудиовыход объединены в единую обучаемую модель. Преимущества данного подхода заключаются в том, что интеграция всех этапов в единую архитектуру минимизирует промежуточные представления и снижает риск потери информации между модулями, что позволяет ускорить процесс синтеза благодаря параллельной обработке, а также упростить настройку и оптимизацию.

Недостатками End-to-End является высокая вычислительная сложность и значительные требования к памяти. Кроме того, несмотря на высокую точность выравнивания, некоторые тонкие нюансы просодии и интонации остаются недостаточно точно смоделированными, что особенно критично при синтезе эмоционально насыщенной речи. Подобные недостатки требуют дальнейшей оптимизации и разработки дополнительных модулей для управления просодическими характеристиками [23].

Non-Autoregressive (NAR) архитектура для синтеза речи представляет собой подход, в котором генерация акустических параметров происходит параллельно для всей последовательности, без традиционного поэтапного авторегрессионного предсказания. В этой модели входной текст сначала преобразуется в последовательность фонем, которая затем конвертируется в векторное представление. Преимущества данной архитектуры заключаются в значительном ускорении процесса синтеза, поскольку модель генерирует всю последовательность мел-спектрограмм одновременно, а также в экономии вычислительных ресурсов при использовании относительно небольшого объёма данных. Явное

моделирование длительности позволяет добиться более точного выравнивания текста с акустическими характеристиками, что приводит к естественному звучанию синтезируемой речи. Однако существуют и ограничения: «обучение NAR моделей требует значительных вычислительных усилий из-за сложности предсказания длительности, а также наблюдаются ограничения в управлении тонкими нюансами просодии, что может приводить к снижению эмоциональной выразительности синтезируемой речи» [37].

Метод DRSpeech (Degradation-Robust Speech) представляет собой подход к синтезу речи, направленный на повышение её устойчивости к различным акустическим искажениям, таким как шум, реверберация и прочие артефакты, возникающие при записи и обработке звуковых данных. Основная идея заключается в использовании многоуровневой обработки входных сигналов, позволяющей извлекать чистые акустические характеристики, минимизируя влияние внешних факторов. Входные данные в DRSpeech включают как текстовую информацию, так и акустические сигналы, содержащие искажения. Для преобразования текста в речь система использует фонемный энкодер, который кодирует входную последовательность в скрытое представление.

Основным преимуществом DRSpeech является его высокая устойчивость к акустическим искажениям, что делает метод применимым в условиях реального мира, например, для синтеза речи в зашумлённых средах или на устройствах с ограниченными микрофонными характеристиками. Также система демонстрирует высокую точность синтеза за счёт эффективного разделения акустических характеристик и фонового шума. Однако метод имеет и ограничения: «высокая вычислительная сложность делает его менее эффективным для задач реального времени, а передача тонких просодических нюансов всё ещё остаётся вызовом, особенно в условиях многоголосового или многоязычного синтеза» [34].

1.4 Обзор технологий VC синтеза речи

Современные методы преобразования голоса (VC) основаны на глубоком обучении и позволяют изменять голосовые характеристики говорящего, сохраняя при этом лингвистическое содержание речи. Метод HiFi-GAN (High-Fidelity Generative Adversarial Network) представляет собой генеративно-состязательный подход к преобразованию голоса, обеспечивающий высокую точность и естественность синтезируемой речи. Основная идея HiFi-GAN заключается в эффективном моделировании акустических паттернов речи с высокой временной и частотной разрешающей способностью. В отличие от традиционных вокодеров, которые могут терять детализацию или требовать больших вычислительных ресурсов, HiFi-GAN использует многорецептивную стратегию обработки аудиосигнала, позволяя получать аудио, практически неотличимое от реальной речи.

Оптимизация модели основана на сочетании противоборственного обучения и спектральных функций потерь. HiFi-GAN может работать на устройствах с ограниченными аппаратными возможностями, демонстрируя высокое качество звука даже при небольшом количестве параметров модели. Однако метод имеет и свои ограничения. Несмотря на достижение высокого качества синтеза, передача тонких просодических нюансов и интонации всё ещё остаётся сложной задачей. Кроме того, «процесс обучения модели требует больших вычислительных ресурсов, а настройка мультипериодических и мультимасштабных дискриминаторов может быть трудоёмкой» [17].

Метод DNN (Deep Neural Network) для преобразования эмоционального голоса представляет собой архитектуру глубокого обучения, предназначенную для точного преобразования спектральных характеристик речи с учетом эмоциональной окраски. В данной модели обработка голосового сигнала основывается на извлечении и последующем преобразовании акустических

признаков. Использование глубоких нейронных сетей позволяет выявлять сложные нелинейные зависимости между входными и выходными параметрами, что обеспечивает более точное преобразование голосовых характеристик.

Принцип работы DNN включает несколько этапов. На первом этапе происходит экстракция признаков, в ходе которой из входного аудиосигнала извлекаются спектральные характеристики, описывающие основные свойства голоса. Затем эти параметры подаются в многослойную нейронную сеть, состоящую из нескольких скрытых слоев, каждый из которых обучается выявлять и моделировать сложные зависимости между акустическими характеристиками. В процессе обработки сеть трансформирует входные параметры в представления, соответствующие целевому голосовому стилю, после чего на выходе генерируются преобразованные акустические характеристики, необходимые для синтеза эмоциональной речи.

Однако данный подход имеет и ряд ограничений. Высокая вычислительная сложность обучения требует значительных аппаратных ресурсов, особенно при работе с большими наборами данных. Кроме того, интерпретируемость моделей глубокого обучения остаётся сложной задачей, поскольку DNN представляет собой "чёрный ящик", затрудняющий анализ принимаемых решений. Также, «несмотря на высокую точность воссоздания основных акустических параметров, передача тонких нюансов просодии и интонации может быть недостаточно точной, что может снижать естественность преобразованной речи» [31].

Метод непарного преобразования голоса представляет собой подход, позволяющий изменять голосовые характеристики говорящего без необходимости в заранее подготовленных парных данных, содержащих идентичные фразы, произнесенные разными людьми. Это делает технологию более гибкой и применимой в условиях ограниченного набора данных, где выровненные пары отсутствуют. Основная идея метода заключается в обучении модели на

несогласованных данных, что требует эффективных механизмов извлечения и преобразования характеристик голоса.

Принцип работы метода основан на трех ключевых этапах. Первым шагом является извлечение характеристик речи, где анализируются как параметры, зависящие от говорящего, так и параметры, отражающие общий лингвистический контент. Далее следует этап отображения, в котором применяется глубокое обучение, включая вариационные автоэнкодеры и генеративные состязательные сети. Эти методы позволяют находить скрытые представления речи, которые могут быть преобразованы в акустические характеристики целевого говорящего. Заключительный этап - реконструкция преобразованной речи, где декодеры используют полученные параметры для генерации итогового аудиовыхода с голосовыми характеристиками целевого говорящего.

При этом данный подход сталкивается с рядом ограничений. «Главной сложностью является точное соответствие голосовых характеристик, так как отсутствие выровненных данных затрудняет процесс адаптации. Наблюдаются проблемы с передачей просодических характеристик, что может снижать естественность синтезированной речи. Обучение таких моделей требует значительных ресурсов, особенно при использовании сложных архитектур» [1].

Архитектура CNN-LSTM представляет собой гибридный подход к обработке речевых сигналов. Принцип работы модели основывается на последовательной обработке входных данных. На первом этапе речевой сигнал представляется в виде мел-спектрограммы, далее свёрточные слои CNN анализируют эти спектрограммы. Каждый сверточный слой применяет фильтры, чтобы выделить значимые элементы спектра. После извлечения признаков результат передается в слои LSTM, которые анализируют временные зависимости и передают информацию о контексте речи. Это позволяет учитывать долговременные связи между элементами входного сигнала, что особенно важно для корректного преобразования интонации и ритма речи в VC-системах.

Преимущества CNN-LSTM заключаются в эффективном выделении спектральных признаков речи и учитывание их временных зависимостей. Гибкость архитектуры CNN-LSTM делает её применимой для обработки различных языков и стилей речи, но данный подход имеет и ряд недостатков. Высокая вычислительная сложность требует значительных ресурсов для обучения и инференса, что ограничивает использование модели в задачах реального времени. Кроме того, хотя архитектура хорошо справляется с общими интонационными характеристиками, она может не всегда точно воспроизводить тонкие просодические нюансы, что влияет на естественность преобразованной речи. Также «для получения высококачественных результатов требуется большое количество размеченных данных, а их недостаток может снижать эффективность обучения» [45].

Метод speaker-independent pre-training представляет собой подход к преобразованию голоса (VC), направленный на устранение зависимости модели от конкретных голосов в обучающем наборе данных. В отличие от традиционных методов VC, требующих парных данных для обучения, этот метод позволяет системе адаптироваться к различным голосам без необходимости в индивидуальной настройке для каждого говорящего.

Ключевыми преимуществами метода являются высокая скорость обучения, возможность работы с неразмеченными данными и улучшенное качество преобразования голоса. В отличие от традиционных методов, требующих выровненных данных для обучения, speaker-independent pre-training упрощает процесс подготовки данных и делает возможной универсальную адаптацию модели к различным голосам. При этом, несмотря на способность воспроизводить основные характеристики голоса, возможны потери тонких просодических нюансов, что может снижать естественность преобразованной речи. «Качество результата зависит от характеристик исходных данных, и если они содержат значительное количество шумов или артефактов, это может негативно сказаться на точности преобразования» [15].

Метод self-supervised learning представляет собой подход к обучению моделей синтеза речи, позволяющий извлекать полезные представления из неразмеченных данных. В TTS-системах SSL применяется для обучения моделей на больших объемах аудиоданных без необходимости ручной разметки, что значительно расширяет возможности адаптации и улучшает качество синтезированной речи. Современные архитектуры используют SSL для автоматического выделения важных фонетических и просодических характеристик речи, что делает их перспективными для интеграции в TTS-модели.

Принцип работы SSL в контексте синтеза речи основан на предварительном обучении модели на неразмеченных аудиоданных с последующей передачей извлеченных представлений. Также SSL часто используется в неавторегрессионных моделях, что повышает скорость работы системы без ущерба для качества. Дополнительным преимуществом является высокая обобщающая способность, позволяющая адаптироваться к различным языкам и голосам без необходимости обучения с нуля. Однако у подхода есть и недостатки. SSL-модели требуют значительных вычислительных ресурсов на этапе обучения, что может ограничивать их применение в условиях ограниченной инфраструктуры. Так же существуют трудности с точной передачей просодических нюансов, что может снизить естественность синтезированной речи. Кроме того, интеграция SSL в существующие TTS-архитектуры увеличивает сложность системы, усложняя процесс отладки и настройки [40].

1.5 Обоснование выбора объекта и предмета исследования

Актуальность проблемы повышения качества, адаптивности и естественности современных систем синтеза речи определяет выбор объекта и предмета исследования. Современные технологии TTS и VC, несмотря на

значительный прогресс, по-прежнему сталкиваются с рядом ограничений, таких как высокая вычислительная сложность, потеря информации между этапами обработки и ограниченная способность к адаптации в условиях недостатка данных. Эти особенности негативно сказываются на итоговом качестве синтезируемой речи и препятствуют широкому практическому применению технологий голосового взаимодействия.

Объект исследования - совокупность процессов синтеза речи. Объект исследования не ограничивается одним конкретным алгоритмом или технологией. Под «совокупностью процессов синтеза речи» понимается весь комплекс этапов, задействованных в создании синтетической речи, что включает:

- Преобразование текста в акустические представления (TTS): этап где текст обрабатывается и преобразуется в промежуточные сигналы, например, в мел-спектрограммы, которые служат основой для дальнейшей генерации аудиосигнала.
- Преобразование голосовых характеристик (VC): процесс включает изменение или адаптацию параметров голоса, что позволяет, например, адаптировать речь одного говорящего к особенностям другого, сохраняя при этом смысловое содержание.
- Промежуточные этапы обработки: включают декодирование, вокодирование и оптимизацию просодических характеристик, обеспечивая корректное преобразование извлечённых признаков в финальный звуковой сигнал.

Предмет исследования - анализ и оценка интеграционных стратегий для объединения алгоритмических подходов TTS и VC в единую модель синтеза речи.

Под интеграционными стратегиями понимается исследование существующих методик и механизмов, направленных на обеспечение эффективного взаимодействия между разрозненными этапами синтеза речи. В рамках этого предмета особое внимание уделяется следующим аспектам:

- Анализ принципов унифицированного представления данных, обеспечивающих беспрепятственную передачу информации между этапами обработки и минимизацию потерь акустических и просодических характеристик.
- Анализ стратегий объединения разрозненных алгоритмов, таких как методы извлечения признаков, генерация акустических представлений и вокодирование, с целью формирования согласованной системы синтеза речи.
- Оценка возможностей использования передовых методов обучения self-supervised learning, для повышения устойчивости и адаптивности модели в условиях ограниченных объемов размеченных данных.
- Исследование алгоритмических подходов, позволяющих перераспределить вычислительную нагрузку между модулями, что способствует снижению эксплуатационных затрат и ускорению синтеза речи.
- Анализ гибридных методов управления тонкими нюансами интонации, ритма и эмоциональной окраски, направленных на обеспечение естественности и выразительности синтезируемого аудиовыхода.

Цель исследования заключается в разработке теоретической концепции интегрированной модульной модели синтеза речи, которая объединяет алгоритмические подходы TTS и VC в единую систему. Для достижения этой цели необходимо решить следующие задачи:

- Исследовать алгоритмические методы, применяемые в современных системах TTS и VC, и определить их основные преимущества и ограничения. Для этого проводится аналитический обзор литературных источников, включающий как теоретический анализ, так и сравнительный анализ существующих моделей.

- Систематизировать проблемы, возникающие при разрозненном применении методов синтеза речи, и выявить ключевые вопросы, требующие решения, что позволит обосновать необходимость интеграции подходов для обеспечения непрерывной передачи информации между этапами обработки.
- Обосновать потенциал интегрированного подхода для повышения качества, адаптивности и эмоциональной выразительности синтезируемой речи. Здесь рассматриваются возможности применения self-supervised learning, а также оценка преимуществ общего промежуточного представления для модульной системы.
- Разработать теоретическую концепцию интеграции алгоритмических методов TTS и VC в единую модель, обеспечивающую устойчивую передачу данных между этапами обработки. Предполагается использование аналитических методов, BPMN-моделирования и экспериментальных данных из литературного обзора.
- Сформулировать рекомендации для дальнейших исследований и практической реализации разработанных алгоритмических решений. Рекомендации будут базироваться на сравнительном анализе существующих методов и выявленных проблемах, что позволит определить перспективные направления оптимизации.

Проведённый аналитический обзор современного состояния технологий синтеза речи показал, что переход от традиционных методов (конкатенативного и статистического параметрического синтеза) к нейронным архитектурам открыл новые возможности для повышения качества, адаптивности и естественности генерируемой речи. Современные системы TTS и VC, использующие методы Sequence-to-Sequence, неавторегрессионные архитектуры, End-to-End подходы, а также методы глубокого обучения, такие как GAN, Autoencoder, DNN и LSTM

демонстрируют значительный прогресс. Однако, несмотря на достигнутые успехи, разрозненное применение этих методов приводит к ряду ограничений:

- Потеря информации между этапами обработки вызывает отсутствие единого промежуточного представления между этапами (извлечение признаков, генерация акустических представлений и преобразование их в аудиосигнал), что способствует несовместимости данных и снижению итогового качества синтеза.
- Ограниченная адаптация к новым данным приводит к высокой зависимости от больших объёмов размеченных данных, что ограничивает возможности персонализированного синтеза речи, особенно при наличии ограниченной обучающей выборки.
- Высокие вычислительные затраты современных архитектур требуют значительных ресурсов на этапах обучения и генерации, что затрудняет их использование в режиме реального времени или на мобильных устройствах.
- Недостаточная управляемость интонацией и просодией не позволяет точно контролировать тонкие нюансы интонации, ритма и эмоциональной окраски синтезируемой речи.
- Раздельное применение технологий TTS и VC приводит к необходимости дублирования процессов и увеличивает вычислительную сложность, снижая эффективность систем в целом.

Анализ литературных источников по современным алгоритмическим подходам в синтезе речи показал, что, хотя отдельные методы TTS и VC демонстрируют значительный прогресс, их разрозненное применение приводит к ряду ограничений. Вместо того чтобы создавать радикально новую модель, возможно применение комбинированной модели, объединяющей сильные стороны как технологий TTS, так и VC. Обоснование предложенного авторского подхода к решению проблемы комбинированности в синтезе речи заключается в следующем:

– Использование единого представления для передачи информации между модулями обеспечивает непрерывность и корректность обработки данных, что существенно снижает потери информации между этапами синтеза речи.
– Применение self-supervised learning позволяет адаптировать модель к новым условиям без необходимости в больших размеченных датасетах. Эти методы способствуют переносу знаний между доменами и автоматическому извлечению полезных признаков из неразмеченных данных, повышая устойчивость и адаптивность системы.
– Интеграция модульного подхода с перераспределением вычислительной нагрузки между этапами обработки снижает эксплуатационные затраты и позволяет ускорить процесс синтеза, что особенно важно для задач реального времени.
– Применение комбинированных методов, объединяющих нейросетевые подходы и вариационные техники, обеспечивает более точное управление тонкими нюансами интонации, ритма и эмоциональной окраски синтезируемой речи, что способствует повышению естественности аудиовыхода.

Предлагаемая интегрированная модель имеет потенциал объединить сильные стороны технологий TTS и VC для решения задач персонализации, мультиязычности и адаптивности в современных голосовых приложениях.

Такая модель может позволить не только компенсировать ограничения отдельных алгоритмов, но и оптимизировать вычислительные процессы за счёт эффективного распределения нагрузки между модулями, а также улучшить контроль над просодическими характеристиками, обеспечивая естественное и эмоционально насыщенное звучание синтезируемой речи.

2 Сравнительный анализ алгоритмов синтеза речи

Данный раздел посвящён сравнительному анализу современных алгоритмических методов синтеза речи, реализующих технологии TTS и VC. Основное внимание уделяется оценке ключевых параметров, таких как точность синтеза, вычислительные затраты, устойчивость к ошибкам и интегрируемость. Проведённый анализ позволяет выявить сильные и слабые стороны разрозненных подходов, определить основные проблемы (например, потерю информации между этапами обработки и высокую вычислительную сложность) и обосновать необходимость интеграции технологий TTS и VC в единую модульную архитектуру. В рамках раздела также рассматриваются перспективы применения self-supervised learning для повышения адаптивности и устойчивости системы. Полученные результаты служат теоретической основой для формирования концептуальной модели, способной удовлетворить современные требования к качеству синтеза речи в условиях реального времени.

2.1 Критерии и метрики сравнения алгоритмов

Сравнительный анализ алгоритмов синтеза речи является необходимым этапом для определения оптимальных решений, способных обеспечить высокое качество, вычислительную эффективность, устойчивость к ошибкам и адаптивность в условиях реального применения. Различные алгоритмические подходы, реализующие технологии TTS и VC, обладают своими сильными и слабыми сторонами, что отражается в их точности синтеза, потреблении вычислительных ресурсов, способности работать с шумовыми входными данными и адаптироваться к изменяющимся условиям эксплуатации.

Для объективной оценки эффективности алгоритмов синтеза речи, реализующих технологии TTS и VC, необходимо определить универсальные

критерии, отражающие ключевые аспекты их практического применения. Для сравнительного анализа были выбраны следующие критерии: точность синтеза, вычислительные затраты, устойчивость к ошибкам и интегрируемость. Выбор данных критериев продиктован необходимостью идентификации оптимальных методов синтеза речи (TTS) и преобразования голоса (VC) для формирования интегрированной системы. Эти критерии позволяют объективно оценить, насколько каждый алгоритм способен обеспечить высокое качество генерируемого аудиовыхода, оптимальное использование вычислительных ресурсов и устойчивую работу системы в реальных условиях эксплуатации. Кроме того, они важны для определения совместимости методов, что критически необходимо для успешного объединения алгоритмических решений в единую модульную архитектуру, способную корректно передавать промежуточные представления между этапами обработки.

Качество синтезируемой речи определяется способностью алгоритма воспроизводить естественные акустические и просодические характеристики. Критерий точность синтеза является ключевым, поскольку конечное качество аудиовыхода непосредственно влияет на восприятие системы пользователями. Для объективной оценки точности синтеза применены следующие метрики:

- MOS (Mean Opinion Score) - позволяет количественно оценить субъективное восприятие естественности и разборчивости речи.
- PESQ (Perceptual Evaluation of Speech Quality) и STOI (Short-Time Objective Intelligibility) - используются для объективной оценки разборчивости и точности воспроизведения аудиосигнала.

Эффективное использование вычислительных ресурсов является важным для применения алгоритмов синтеза речи. Основными показателями вычислительных затрат являются:

- Время инференса в секундах, позволяющее оценить скорость генерации аудиовыхода.

– Объём потребляемой оперативной памяти и вычислительной мощности, характеризующий ресурсозатратность алгоритма.

В реальных условиях входные данные могут быть зашумлёнными или неполными, что требует от алгоритмов способности корректно обрабатывать ошибки на этапах предварительной обработки. Для оценки устойчивости к ошибкам, применены следующие метрики процентного снижения показателей качества (MOS, PESQ, STOI) при внесении контролируемых искажений в входной сигнал.

Интегрируемость оценивает возможность успешного объединения алгоритмических решений в единую модульную архитектуру, что является ключевым для обеспечения непрерывной передачи промежуточных представлений между компонентами системы. Этот критерий предусматривает следующие аспекты:

– Выбранные методы TTS и VC должны использовать единое представление в виде мел-спектрограмм, которое позволяет выводу одного модуля служить корректным входом для следующегоя, без необходимости дополнительных преобразований.
– Оценка интегрируемости включает анализ временных задержек (latency) при передаче данных между модулями и степень потерь информации. Степень совпадения промежуточных представлений оценивается с использованием мер сходства (например, косинусного сходства).
– Применение self-supervised learning должно способствовать улучшению взаимодействия между компонентами, позволяя динамично обновлять модель без нарушения общего потока данных, что обеспечит более слаженную работу системы за счёт эффективного переноса знаний между модулями TTS и VC.

2.2 Сравнительный анализ TTS моделей.

Изолированное тестирование отдельных алгоритмов зачастую не отражает реальную эффективность, поскольку компоненты системы работают в комплексной среде, где взаимодействие между модулями существенно влияет на итоговое качество как синтеза, так и преобразования голоса. Поэтому анализ ведётся на основе интегрированных моделей, каждая из которых использует стандартное решение базового алгоритма. Результаты сравнительного анализа моделей TTS представлены в таблицах 1-5.

Следует отметить, что приведённые в таблицах значения являются ориентировочными оценками, полученными на основе анализа общих тенденций и сравнительных данных из публикаций [41], а также на основе эмпирических результатов, представленных исследованиях по методам TTS. Эти показатели представляют собой обобщённые средние и могут варьироваться в зависимости от конкретных экспериментальных условий. Эмпирическая проверка данных результатов не входит в рамки настоящей магистерской диссертации, но представляет собой важное и перспективное направление для будущих исследований.

Таблица 1 - Анализ модели Tacotron 2

Критерий	Показатель	Обоснование
Точность синтеза	MOS: 4.0-4.5 PESQ: 2.5-3.0 STOI: 0.85-0.90	Tacotron 2 обеспечивает высокое качество синтезированной речи за счёт эффективного end-to-end обучения и использования мел-спектрограмм, что подтверждается высокими субъективными (MOS) и объективными (PESQ, STOI) оценками.

Продолжение таблицы 1

Критерий	Показатель	Обоснование
Вычислительные затраты	Время инференса: 0.3-0.8 с/ед. синтезированной речи Объём памяти: 2-4 ГБ	Авторегрессивная генерация Tacotron 2 требует последовательного вычисления каждого шага, что увеличивает время инференса по сравнению с неавторегрессивными моделями. Дополнительно, RNN-based модели требуют больше оперативной памяти.
Устойчивость к ошибкам	Снижение MOS, PESQ, STOI на 10-20% при внесении контролируемых искажений	Из-за чувствительности к ошибкам на входных данных из-за, шумов и неполноты сигнала, Tacotron 2 может демонстрировать снижение качества синтеза, что подтверждается уменьшением основных показателей.
Интегрируемость	Использование мел-спектрограмм обеспечивает легкую интеграцию; возможны задержки при передаче данных между модулями	Применение стандартного представления в виде мел-спектрограмм позволяет без дополнительных преобразований передавать данные между модулями TTS и VC. Однако авторегрессивная природа модели может вызвать задержки, а применение self-supervised learning способно улучшить совместимость.

Общий анализ показывает, что Tacotron 2 генерирует речь высокого качества, что подтверждается хорошими значениями метрик. Однако авторегрессивный характер модели приводит к увеличенным затратам вычислительных ресурсов, выраженным в длительном времени инференса и повышенном потреблении памяти. Кроме того, Tacotron 2 демонстрирует чувствительность к искажениям входных данных, что может негативно сказываться на стабильности аудиовыхода в условиях реального использования. Важно отметить, что использование мел-спектрограмм в качестве единого представления способствует эффективной интеграции компонентов TTS и VC, обеспечивая корректную передачу промежуточных представлений между модулями. Применение self-supervised learning может дополнительно улучшить согласованность работы системы в условиях изменяющихся эксплуатационных параметров.

Таблица 2 - Анализ модели FastSpeech 2

Критерий	Показатель	Обоснование
Точность синтеза	MOS: 4.2-4.6 PESQ: 2.8-3.3 STOI: 0.90-0.95	FastSpeech 2 достигает высоких значений показателей качества синтеза, что обусловлено использованием целевых мел-спектрограмм и эффективными механизмами контроля длительности фонем. Такой подход позволяет воспроизводить тонкие просодические и акустические нюансы речи, что подтверждается данными из рецензируемых источников по TTS.
Вычислительные затраты	Время инференса: 0.01-0.05 с на единицу синтезируемой речи Потребление памяти: 1-2 ГБ	Feed-forward Transformer архитектура FastSpeech 2 обеспечивает параллельную генерацию мел-спектрограмм, что значительно сокращает время инференса. Потребление оперативной памяти также снижено за счёт оптимизированного процесса генерации, что делает модель эффективной для развертывания в системах реального времени.
Устойчивость к ошибкам	Снижение показателей (MOS, PESQ, STOI) на 5-15% при внесении контролируемых искажений	Благодаря отсутствию механизма внимания, характерного для авторегрессивных моделей, FastSpeech 2 менее подвержена ошибкам, таким как пропуски или повторения. Это обеспечивает более стабильное качество синтеза даже при наличии искусственно внесённых шумовых искажений, демонстрируя высокую устойчивость к ошибкам ввода.
Интегрируемость	Использование мел-спектрограмм в качестве представления, что обеспечивает минимальные задержки при передаче данных	Унифицированное представление в виде мел-спектрограмм позволяет без дополнительных преобразований передавать данные между компонентами системы синтеза речи. Параллельная структура модели способствует снижению временных задержек.

Проведённый сравнительный анализ показывает, что FastSpeech 2 является высокоперспективным решением в области синтеза речи. Модель достигает конкурентоспособного качества аудиовыхода. Благодаря своей параллельной feed-

forward архитектуре FastSpeech 2 демонстрирует значительно сокращённое время инференса и сниженные требования к оперативной памяти, что положительно сказывается на вычислительной эффективности системы синтеза речи.

Таблица 3 - Анализ модели DurIAN

Критерий	Показатель	Обоснование
Точность синтеза	MOS: 4.0-4.4 PESQ: 2.7-3.2 STOI: 0.87-0.92	DurIAN демонстрирует высокое качество синтеза, засчет использования duration informed attention network для точного моделирования длительности фонем, что позволяет лучше воспроизводить просодические и акустические характеристики речи.
Вычислительные затраты	Время инференса: 0.05-0.15 с на единицу синтезируемой речи Потребление памяти: 1.5-2.5 ГБ	feed-forward архитектура и использование duration predictor значительно сокращается время генерации по сравнению с авторегрессивными моделями. При этом ресурсы, необходимые для обработки, остаются на относительно низком уровне, что делает модель пригодной для интеграции в системы реального времени.
Устойчивость к ошибкам	Снижение показателей (MOS, PESQ, STOI) на 8-18% при внесении контролируемых искажений	Применение duration informed attention способствует более надёжной привязке входных данных к сгенерированному аудио, что уменьшает вероятность ошибок. Это повышает устойчивость модели к шумовым воздействиям, что особенно важно при использовании в сложных акустических условиях.
Интегрируемость	Использование мел-спектрограмм для представления аудиоданных; низкие временные задержки при передаче информации между модулями	Унифицированное представление данных в виде мел-спектрограмм облегчает обмен информацией между компонентами интегрированной системы синтеза речи. Feed-forward структура обеспечивает минимальные задержки.

Анализ характеристик DurIAN свидетельствует о том, что модель способна генерировать речь высокого качества, что подтверждается ориентировочными

значениями MOS, PESQ и STOI, демонствующими хорошее воспроизведение просодических особенностей. Благодаря feed-forward архитектуре и использованию duration predictor DurIAN демонстрирует существенно сокращённое время инференса и относительно низкое потребление оперативной памяти, что делает её привлекательной для систем реального времени. Также применение duration informed attention network позволяет модели эффективно связывать входные данные с аудиовыходом, что повышает устойчивость при наличии шумовых искажений. Унифицированное представление аудиоданных в виде мел-спектрограмм облегчает интеграцию DurIAN с другими компонентами модульной системы синтеза речи, а перспективное применение self-supervised learning обещает дальнейшее улучшение взаимодействия между модулями.

Таблица 4 - Анализ модели Tacotron 1

Критерий	Показатель	Обоснование
Точность синтеза	MOS: 3.5-4.0 PESQ: 2.0-2.5 STOI: 0.80-0.85	Обеспечивает удовлетворительное качество синтеза, однако результаты уступают более современным моделям. Умеренное качество аудиовыхода, обусловленное ограничениями архитектуры.
Вычислительные затраты	Время инференса: 1.0-2.0 с на единицу синтезируемой речи Потребление памяти: 1-2 ГБ	Архитектура Tacotron 1 требует последовательного выполнения каждого шага, что приводит к значительному увеличению времени генерации. Требования к памяти могут быть несколько ниже, чем у современных моделей.
Устойчивость к ошибкам	Снижение показателей (MOS, PESQ, STOI) на 15-25% при внесении контролируемых искажений	За счёт использования механизма внимания модель Tacotron 1 может сталкиваться с проблемами синхронизации между текстом и аудиовыходом, что приводит к более выраженному снижению качества при наличии шумовых искажений.
Интегрируемость	Использует линейные спектрограммы для представления аудиовыхода;	Tacotron 1 генерирует линейные спектрограммы. Авторегрессивная природа модели создаёт задержки при передаче данных между модулями, что влияет на общую интегрируемость.

Анализ показал, что Tacotron 1, как одна из ранних нейронных TTS-моделей, демонстрирует удовлетворительное, но в целом уступающее качество синтеза по сравнению с современными решениями. Авторегрессивная архитектура приводит к значительному времени инференса и повышенным вычислительным затратам, что делает её менее пригодной для реальных систем, работающих в режиме реального времени. Дополнительно модель характеризуется чувствительностью к шуму, что проявляется в существенном снижении показателей качества при внесении искажений. При этом, специфика генерируемых линейных спектрограмм требует дополнительных этапов преобразования для интеграции с компонентами, использующими мел-спектрограммы, что усложняет её объединение в единую модульную архитектуру синтеза речи.

Таблица 5 - Итоговая таблица сравнительного анализа моделей TTS

Модель	Точность синтеза	Вычислительные затраты	Устойчивость к ошибкам	Интегрируемость
Tacotron 2	✓	-	-	✓
FastSpeech 2	✓	✓	✓	✓
DurIAN	✓	✓	✓	✓
Tacotron 1	-	✗	✗	-

Общий сравнительный анализ алгоритмов синтеза речи показал, что эффективность современных методов значительно различается по таким ключевым параметрам, как точность синтеза, вычислительные затраты, устойчивость к ошибкам и интегрируемость. Из рассмотренных категорий алгоритмов видно, что решения, основанные на последовательной авторегрессивной генерации (End-to-End и Seq2Seq модели, представленные Tacotron 2 и Tacotron 1), несмотря на высокий показатель качества аудиовыхода, страдают от значительных вычислительных затрат и повышенной чувствительности к шуму. Это ограничивает их практическое применение в условиях реального времени, а также затрудняет их

интеграцию в единую модульную архитектуру за счёт несовершенной передачи промежуточных представлений.

В противоположность им, подходы, реализованные в моделях Non-autoregressive и Robust TTS решений, демонстрируют более высокую вычислительную эффективность, обеспечивая короткое время инференса и умеренное потребление ресурсов. Благодаря параллельной генерации мел-спектрограмм, FastSpeech 2 минимизирует задержки, а отказ от сложных механизмов внимания снижает вероятность ошибок, что положительно сказывается на стабильности синтеза. DurIAN, в свою очередь, за счёт использования duration informed attention network успешно моделирует длительность фонем, что улучшает просодию и устойчивость модели к шумовым воздействиям. Унифицированное представление аудиоданных в виде мел-спектрограмм способствует безпроблемной интеграции этих алгоритмов с другими компонентами системы, а потенциал их комбинирования self-supervised learning обеспечивает возможность динамической адаптации при изменяющихся условиях эксплуатации.

Анализ показал, что современные алгоритмы синтеза речи, реализованные на базе Non-autoregressive архитектур и Robust TTS решений, демонстрируют явные преимущества по сравнению с традиционными методами. Оба подхода характеризуются высокой вычислительной эффективностью за счёт использования feed-forward архитектур, что обеспечивает параллельную генерацию мел-спектрограмм и, соответственно, сокращённое время инференса и умеренное потребление вычислительных ресурсов.

При этом каждое из этих решений имеет свои алгоритмические особенности. FastSpeech 2 использует параллельную генерацию мел-спектрограмм посредством длиномерного регулятора, что минимизирует задержки и делает модель менее чувствительной к ошибкам, типичным для механизмов внимания, часто встречающихся в авторегрессивных системах. Такой подход обеспечивает высокую

скорость и стабильность работы, что критически важно для задач в реальном времени и для эффективной интеграции с другими компонентами.

С другой стороны, DurIAN, как представитель Robust TTS решений, применяет duration informed attention network, что позволяет точно моделировать длительность фонем и воспроизводить тонкие просодические и акустические нюансы. Это даёт модели явное преимущество в устойчивости к шумовым воздействиям и точности аудиовыхода, хотя может несколько усложнять процесс оптимизации по сравнению с полностью неавторегрессивными архитектурами.

Для формирования единой модульной системы критически важны не только высокое качество синтеза, но и способность алгоритма корректно передавать промежуточные представления между модулями TTS и VC с минимальными задержками. В этом контексте, хотя оба подхода демонстрируют сильные стороны, предпочтение следует отдавать решению, минимизирующему затраты времени и ресурсов и обеспечивающему более простое взаимодействие между компонентами.

При учёте специфики интеграции, где ключевыми параметрами являются скорость синтеза, экономия вычислительных ресурсов и минимальные задержки при передаче данных между модулями, оптимальным решением для формирования интегрированной модульной системы синтеза речи представляется алгоритмический подход, реализованный в моделях на базе non-autoregressive архитектуры с параллельной feed-forward структурой. Такой подход обеспечивает значительное сокращение времени инференса и сниженные вычислительные затраты, что критически важно для систем, работающих в режиме реального времени. Отсутствие сложных механизмов внимания делает эти модели менее чувствительными к ошибкам, характерным для авторегрессивных решений, а единое представление аудиоданных способствует лёгкой интеграции с другими компонентами, такими как модули преобразования голоса.

В отличие от этого, Robust TTS решения, такие как DurIAN, хотя и демонстрируют высокое качество синтеза за счёт использования duration informed

attention network для точного моделирования длительности фонем, характеризуются вычислительной сложностью и средним временем инференса. Это обусловлено необходимостью реализации дополнительных механизмов для борьбы с потерей информации в условиях искажённых входных данных. Таким образом Robust TTS решения не могут обеспечить столь же низкие вычислительные затраты и минимальные задержки, как модели, основанные на non-autoregressive архитектуре.

2.3 Сравниетельный анализ VC моделей.

Приведённые в таблицах значения являются ориентировочными оценками, полученными на основе анализа общих тенденций и сравнительных данных из публикаций [44], [3], [22], [14], [30].

При проведении сравнительного анализа, результаты которого представлены в таблицах 6-11, были рассмотрены следующие модели VC, каждая из которых иллюстрирует одну из категорий алгоритмических подходов:

- DNN-основанные модели: SIEVC - модель, основанная на глубоких нейронных сетях для преобразования голосовых характеристик с сохранением лингвистического содержания. «Обучается на параллельных данных, что позволяет модели точно передавать индивидуальные особенности голоса и обеспечивать стабильные результаты» [44].
- GAN-архитектуры: StarGAN-VC++ - это многодоменная модель, использующая генеративно-состязательные сети (GAN) для преобразования голоса, которая направлена на сохранение эмоционального состояния исходного говорящего [3].
- LSTM модели: Deep Bidirectional LSTM - модель, использующая рекуррентные нейронные сети с двунаправленной архитектурой для учета временной динамики речи. «Обработка последовательности в обоих

направлениях позволяет захватывать контекстную информацию, что улучшает передачу интонационных, просодических и эмоциональных нюансов при преобразовании голоса» [22].

- Подходы для обучения на непарных данных: ACVAE-VC - модель, основанная на вариационных автоэнкодерах, которая предназначена для преобразования голоса без необходимости наличия парных данных. «Используя полностью свёрточные кодировщики и декодировщики, модель эффективно улавливает временные зависимости в акустических признаках как исходной, так и целевой речи, что улучшает плавность и качество синтезируемого звука» [14].
- Speaker independent pre training: JSV-VC (Jointly Trained Speaker Verification) модель, которая объединяет задачи преобразования голоса с сохранением эмоциональной окраски в рамках единого этапа предварительного обучения. «Используя огромный объём неразмеченных данных, модель обучается выделять универсальные голосовые признаки и ключевые эмоциональные характеристики, не привязанные к отдельным индивидуальным особенностям говорящего.» [30]

Таблица 6 - анализ модели SIEVC

Критерий	Показатель	Обоснование
Точность синтеза	MOS: 4.2-4.7 PESQ: 2.8-3.2 STOI: 0.86-0.92	Модель SIEVC обеспечивает высокое качество преобразования эмоционального голоса благодаря эффективному разделению информации, позволяющему сохранять естественные интонационные и просодические характеристики.
Вычислительные затраты	Время инференса: 0.4-0.9 с/ед. Синтезируемой речи голоса Объём памяти: 3-5 ГБ	Применение структуры автоэнкодера в сочетании с GAN повышает вычислительную нагрузку, однако улучшение качества аудиосигнала оправдывает эти затраты..

Продолжение таблицы 6

Критерий	Показатель	Обоснование
Устойчивость к ошибкам	Снижение показателей (MOS, PESQ, STOI) на 15-25% при контролируемых искажениях	Применение методов минимизации взаимной информации позволяет модели эффективно справляться с зашумлёнными данными, остается чувствительность к входным искажениям..
Интегрируемость	Использование мел-спектрограмм обеспечивает лёгкую интеграцию между модулями	Унифицированное представление аудиоданных посредством мел-спектрограмм.

Модель SIEVC демонстрирует отличное качество преобразования голоса, что подтверждается высокими показателями качества. Благодаря архитектуре, обеспечивающей раздельную обработку эмоциональной и базовой информации, синтезированная речь получается естественной и разборчивой. С точки зрения вычислительных характеристик, модель демонстрирует быстрое время инференса при умеренных затратах оперативной памяти, что делает её пригодной для использования в системах реального времени. Несмотря на некоторую чувствительность к шумовым воздействиям, SIEVC сохраняет стабильное качество работы.

Таблица 7 - Анализ модели StarGAN-VC

Критерий	Показатель	Обоснование
Точность синтеза	MOS: 4.1-4.3 PESQ: 2.8-3.2 STOI: 0.87-0.92	StarGAN-VC++ демонстрирует высокий уровень натуральности синтеза за счёт эффективного разделения информации о голосе и эмоциях. Высокие показатели MOS, а предполагаемые значения PESQ и STOI указывают на хорошую разборчивость и естественность речи.

Продолжение таблицы 7

Критерий	Показатель	Обоснование
Вычислительные затраты	Время инференса: 0.2-0.6 с на единицу синтезированной речи Объём памяти: 1.5-3 ГБ	Благодаря использованию неавторегрессивного подхода и оптимизированной архитектуре, StarGAN-VC++ способен обеспечить короткое время генерации аудиовыхода. Простота архитектуры способствует снижению расхода вычислительных ресурсов и памяти.
Устойчивость к ошибкам	Снижение показателей (MOS, PESQ, STOI) на 15-25% при контролируемых искажениях	Модель демонстрирует умеренную чувствительность к ошибкам на входе. При внесении контролируемых искажений наблюдается снижение ключевых метрик на 15-25%, что свидетельствует о необходимости дальнейшей оптимизации устойчивости, несмотря на высокое качество синтеза в оптимальных условиях.
Интегрируемость	Использование мел-спектрограмм; низкие задержки при передаче данных	Унифицированное представление аудиоданных в виде мел-спектрограмм позволяет легко передавать промежуточные представления между модулями TTS и VC без дополнительных преобразований.

Анализ характеристик модели StarGAN-VC++ показывает, что она обладает высоким качеством синтеза благодаря точной передаче эмоциональной информации и естественному звучанию преобразованной речи. Применение неавторегрессивной архитектуры обеспечивает достаточно быстрое время инференса и эффективное использование вычислительных ресурсов, что делает её подходящей для систем, работающих в реальном времени. Несмотря на умеренную чувствительность к шуму, модель демонстрирует стабильные результаты при конвертации, что подтверждается высокими значениями MOS и предполагаемой устойчивостью по объективным метрикам. Унифицированное использование мел-спектрограмм способствует лёгкой интеграции с другими компонентами модульной системы.

Таблица 8 - Анализ модели DBLSTM

Критерий	Показатель	Обоснование
Точность синтеза	MOS: 4.0-4.5 PESQ: 2.5-3.0 STOI: 0.85-0.90	Использование концепции непрерывного вейвлет-преобразования обеспечивает высокое качество извлечения акустических и просодических характеристик. Высокие значения STOI свидетельствуют об устойчивости к изменениям и шумам.
Вычислительные затраты	Время инференса: ~0.4 с на единицу синтезированной речи Объём памяти: 3-5 ГБ	Модель на базе DBLSTM, несмотря на высокую эффективность в воспроизведении речи, требует значительных вычислительных ресурсов из-за сложности архитектуры.
Устойчивость к ошибкам	Снижение основных показателей (MOS, PESQ, STOI) на 15-25% при контролируемых искажениях	При внесении контролируемых шумовых искажений наблюдается заметное снижение ключевых показателей качества синтеза. Это связано с чувствительностью модели к входным ошибкам, что указывает на необходимость дополнительной предобработки входных данных.
Интегрируемость	Использование CWT для формирования представления аудиоданных; задержки при передаче данных	Применение непрерывного вейвлет-преобразования позволяет получить единое представление аудиосигнала, что значительно облегчает передачу промежуточных данных между модулями системы .

Модель демонстрирует высокое качество синтеза голоса, что подтверждается стабильными значениями субъективных и объективных показателей. Применение методов непрерывного вейвлет-преобразования для моделирования фундаментальной частоты и энерго-контуров позволяет эффективно воспроизводить просодические особенности речи. Однако архитектура, основанная на двунаправленной LSTM, требует значительных вычислительных ресурсов, что ведёт к относительно длительному времени генерации и увеличенному объёму памяти при работе с большими данными. При этом модель проявляет умеренную

чувствительность к ошибкам: при наличии шумовых искажений наблюдается существенное снижение качества синтеза. Использование унифицированного представления аудиоданных существенно облегчает интеграцию данной модели с другими компонентами системы, обеспечивая минимальные задержки при обмене промежуточными данными.

Таблица 9 - Анализ модели ACVAE-VC

Критерий	Показатель	Обоснование
Точность синтеза	MOS: 3.5-4.0 PESQ: 2.0-2.5 STOI: 0.80-0.85	Модель демонстрирует удовлетворительное качество синтеза речи. Использование вариационного автоэнкодера в полностью сверточной архитектуре позволяет сохранить основные акустические и просодические характеристики.
Вычислительные затраты	Время инференса: 0.2-0.6 с/ед. синтезированной речи Объём памяти: 1.5-3 ГБ	За счёт применения свёрточных слоёв модель обеспечивает параллельную генерацию, что позволяет существенно сократить время синтеза по сравнению с рекуррентными архитектурами.
Устойчивость к ошибкам	Снижение показателей (MOS, PESQ, STOI) на 5-15% при искажениях	Модель демонстрирует определённую чувствительность к шумовым воздействиям и другим искажениям входного сигнала, что приводит к снижению основных показателей качества на 5-15%.
Интегрируемость	Использование мел-спектрограмм; возможны небольшие задержки при передаче данных	Применение единичного представления аудиоданных в виде мел-спектрограмм значительно упрощает объединение выходных данных между компонентами системы.

Модель ACVAE-VC демонстрирует удовлетворительное качество синтеза голоса, что подтверждается стабильными значениями MOS, PESQ и STOI. Использование полностью свёрточной архитектуры позволяет ускорить процесс генерации по сравнению с рекуррентными решениями, при этом требования к оперативной памяти остаются умеренными. Несмотря на это, модель чувствительна

к шумовым искажениям. Унифицированное представление аудиоданных посредством мел-спектрограмм способствует лёгкой интеграции с другими модулями системы. При этом потенциальные задержки при передаче данных между модулями остаются незначительными.

Таблица 10 - Анализи модели SIEVC

Критерий	Показатель	Обоснование
Точность синтеза	MOS: 3.7-3.8 PESQ: 2.5-3.0 STOI: 0.85-0.90	Модель демонстрирует высокое качество эмоциональной конверсии, сохраняя естественные акустические и просодические характеристики. Применение дисентанглированного представления с минимизацией взаимной информации позволяет эффективно разделить эмоциональные признаки и фоновые характеристики.
Вычислительные затраты	Время инференса: 0.3-0.8 с/ед. синтезированной речи Объём памяти: 2-4 ГБ	За счёт оптимизированной сверточной архитектуры с использованием эффективных операций модель обеспечивает параллельную генерацию с минимальными задержками.
Устойчивость к ошибкам	Снижение показателей (MOS, PESQ, STOI) на 10-20% при искажениях	Несмотря на применение методов дисентанглирования для разделения эмоциональных и неэмоциональных признаков, модель сохраняет умеренную чувствительность к шумовым воздействиям и другим искажениям входного сигнала. Это приводит к снижению ключевых метрик качества на 10-20% при добавлении контролируемых искажений.
Интегрируемость	Использование мел-спектрограмм; небольшие задержки при передаче промежуточных данных	Единый формат мел-спектрограмм позволяет напрямую передавать данные между модулями TTS и VC без необходимости дополнительных преобразований.

Эффективное разделение эмоциональных и фоновых характеристик обеспечивает высокое качество синтеза, подтверждаемое стабильными

субъективными и объективными оценками. Оптимизированная сверточная архитектура обеспечивает параллельное выполнение операций, что позволяет поддерживать низкое время инференса и умеренные требования к ресурсам, делая модель применимой в условиях реального времени. Несмотря на выявленную умеренную чувствительность к входным искажениям, методика дисентанглирования признаков позволяет поддерживать приемлемый уровень качества, а использование унифицированного представления в виде мел-спектрограмм облегчает интеграцию модели в комплексную модульную систему.

Таблица 11 - Итоговая таблица сравнительного анализа моделей VC

Модель	Точность синтеза	Вычислительные затраты	Устойчивость к ошибкам	Интегрируемость
SIEVC	✓	-	✓	✓
StarGAN-VC++	-	✓	-	✓
LSTM-VC	✓	-	✓	✓
ACVAE-VC	✓	✓	✓	✓
JSV-VC	✓	-	✓	✓

Общий сравнительный анализ алгоритмов преобразования голоса показал, что эффективность современных методов значительно различается по ключевым параметрам: точности преобразования, вычислительным затратам, устойчивости к ошибкам и интегрируемости. SIEVC и ACVAE, благодаря своей feed-forward структуре, позволяют реализовать параллельную генерацию выходных параметров, что способствует существенному сокращению задержек и снижению нагрузки на вычислительные ресурсы. При этом они сохраняют способность точно моделировать спектральные характеристики и просодические нюансы речи, что обеспечивает стабильное и высококачественное преобразование голоса. Такая вычислительная эффективность и устойчивость к ошибкам делают их особенно

привлекательными для интеграции в модульные системы, где требуется быстрое взаимодействие между компонентами, а также совместимость с self-supervised learning для динамической адаптации системы к изменяющимся условиям эксплуатации.

Можно отметить, что LSTM-VC не является «непригодной» с точки зрения чистого качества преобразования, однако её последовательный характер создаёт дополнительные временные и ресурсные затраты, что особенно ограничивает её применение в системах с жесткими требованиями к быстродействию. В отличие от них, модели, основанные на полностью свёрточных архитектурах, гораздо легче масштабируются и интегрируются в единую систему, поскольку единое промежуточное представление, например, мел-спектрограмм, передаётся между модулями без дополнительных преобразований, что минимизирует временные задержки и обеспечивает слаженную работу комплексной системы преобразования голоса.

Для формирования единой модульной системы преобразования голоса критически важны не только высокое качество конвертации, но и способность алгоритма точно передавать промежуточные представления между компонентами системы с минимальными задержками. В этом контексте, несмотря на то, что различные архитектурные подходы на базе глубоких нейронных сетей, GAN-архитектуры или рекуррентные сети демонстрируют определённые преимущества, предпочтение следует отдавать решениям, которые способны обеспечить быстрое время инференса и низкие вычислительные затраты, сохраняя при этом высокое качество преобразования и устойчивость к шумовым воздействиям.

При учёте специфики интеграции, где ключевыми параметрами являются скорость преобразования, экономия вычислительных ресурсов и минимизация задержек при передаче промежуточных представлений между компонентами системы, оптимальным решением для формирования интегрированной модульной системы преобразования голоса является использование моделей, основанных на

полностью свёрточных архитектурах. Такие модели обеспечивают параллельную генерацию выходных параметров, что существенно сокращает время обработки, позволяя работать в режиме реального времени и эффективно использовать вычислительные ресурсы. Отсутствие рекуррентных слоёв и сложных механизмов внимания снижает чувствительность системы к ошибкам, возникающим при передаче данных, а унифицированное представление аудиоданных в виде мел-спектрограмм обеспечивает лёгкую интеграцию с другими компонентами системы.

Особое внимание в данной архитектурной парадигме уделяется методам, позволяющим работать с непарными данными. В этом контексте модель ACVAE-VC выступает в качестве предпочтительного кандидата, поскольку она способна обучаться на неструктурированных данных, не требуя наличия строго выровненных пар примеров. Это существенно расширяет область её применения, так как позволяет системе адаптироваться к условиям ограниченного набора обучающих примеров, сохраняя при этом высокое качество синтеза и преобразования голосовых характеристик. Кроме того, архитектурное решение ACVAE-VC демонстрирует сбалансированные показатели по времени инференса и вычислительным затратам, что делает её особенно привлекательной в системах, где критически важна скорость обработки данных и минимальные задержки при передаче информации между модулями.

В отличие от моделей, использованных в методах предварительного обучения на больших объёмах данных, SIEVC классифицируется как подход, основанный на глубоких нейронных сетях (DNN). Такие модели демонстрируют высокое качество синтеза за счёт точного моделирования спектральных характеристик и сохранения лингвистического содержания, однако их DNN-архитектура требует значительно большего объёма вычислительных ресурсов и увеличивает время инференса. Это может приводить к дополнительным задержкам при передаче промежуточных представлений между модулями системы. Таким образом, несмотря на высокую

точность синтеза, модели, реализованные как DNN-подход (например, SIEVC), оказываются менее оптимальными с точки зрения вычислительной эффективности и скорости работы, что особенно критично для интегрированных систем, функционирующих в режиме реального времени.

Исходя из представленных аргументов, для формирования единой модульной системы преобразования голоса приоритет отдан моделям, основанным на полностью свёрточных архитектурах. В данной парадигме особое внимание уделяется подходам для обучения на непарных данных, так как они позволяют обучаться без необходимости наличия строго выровненных пар примеров, что существенно расширяет область применения системы при ограниченном объёме обучающих данных. В этом контексте модель ACVAE-VC, основанная на вариационных автоэнкодерах и использующая полностью свёрточные кодировщики и декодировщики для эффективного улавливания временных зависимостей в акустических признаках, демонстрирует сбалансированные показатели по скорости инференса и вычислительным затратам. Такой подход позволяет реализовать параллельную генерацию выходных параметров, что существенно сокращает время обработки, повышает вычислительную эффективность и минимизирует задержки при передаче промежуточных представлений между компонентами системы. При этом отсутствуют рекуррентные слои и сложные механизмы внимания, что снижает чувствительность системы к ошибкам, возникающим в условиях шумовых воздействий. Несмотря на высокую точность синтеза, модели, реализованные по DNN-подходу, оказываются менее оптимальными с точки зрения вычислительной эффективности и скорости работы для систем реального времени. Исходя из этих соображений, предпочтение отдано алгоритмам, основанным на полностью свёрточных архитектурах, в частности, подходу для обучения на непарных данных, представленному в модели ACVAE-VC. Такой выбор обеспечивает оптимальный баланс между качеством преобразования, быстродействием и экономией вычислительных ресурсов, что

является ключевым для работы интегрированной системы преобразования голоса в условиях реального времени и её дальнейшей адаптации посредством self-supervised learning.

Проведённый сравнительный анализ современной литературы и экспериментальных данных позволил выделить ключевые направления развития технологий синтеза речи, реализуемых в рамках систем TTS и VC. Исследование показало, что эффективность рассматриваемых методов существенно различается по таким критически важным параметрам, как точность синтеза, вычислительные затраты, устойчивость к ошибкам, адаптивность и интегрируемость.

На основании полученных результатов, для формирования интегрированной модульной системы синтеза речи выбраны следующие технологии и их модели:

- FastSpeech 2 - оптимальное решение для реализации TTS. Эта модель основана на non-autoregressive технологии, которая позволяет генерировать выходные данные без пошаговой авторегрессии, снижая задержки при синтезе. Благодаря параллельной feed-forward архитектуре модель демонстрирует высокую скорость инференса, низкое потребление вычислительных ресурсов и стабильное качество синтеза.
- ACVAE-VC - в качестве модели для преобразования голоса, поскольку подход на основе обучения на непарных данных демонстрирует сбалансированные показатели по точности преобразования и вычислительной эффективности, что особенно важно при ограниченных объёмах обучающих данных.

Эти технологии и их модели обладают высокой совместимостью между собой и self-supervised learning. Таким образом, выбранные решения обеспечивают корректную передачу промежуточных представлений между отдельными компонентами системы, а также позволяют динамично адаптировать модель к условиям реального времени. Результаты сравнительного анализа служат теоретической основой для дальнейшего формирования и реализации

интегрированной гибридной архитектуры синтеза речи, которая объединит преимущества TTS и VC в единой, универсальной системе.

2.4 Оценка влияния интеграций на качество синтеза

В предлагаемой схеме интеграции используются FastSpeech 2 в качестве TTS-модуля и ACVAE-VC в качестве VC-модуля. Анализ имеющихся данных позволяет сформулировать следующие выводы относительно влияния объединения компонентов на итоговое качество синтеза:

- Использование единого формата мел-спектрограмм для передачи данных между модулями, позволяет сохранить важные акустические и просодические характеристики. Такой подход снижает риск потери информации между этапами и поддерживает естественность звучания.
- Применение self-supervised learning способствует динамической адаптации объединённой системы к изменяющимся языковым и акустическим условиям. Это уменьшает зависимость от объемов размеченных данных и повышает устойчивость системы к шуму и искажениям входного сигнала.
- Объединение модулей с параллельной обработкой позволяет существенно сократить время инференса без снижения параметров качества. При оптимальной передаче промежуточных представлений система сможет демонстрировать приемлемый компромисс между скоростью работы и качеством аудиовыхода.
- Синергия объединённых компонентов позволяет компенсировать отдельные недостатки разрозненных методов. FastSpeech 2, несмотря на свою высокую скорость и вычислительную эффективность, может иногда терять тонкие просодические нюансы из-за ограничений duration predictor, а ACVAE-VC может демонстрировать некоторую нестабильность в преобразовании голосовых характеристик, особенно при ограниченных

объемах обучающих данных. При корректно настроенной синхронизации модулей эти недостатки компенсируются: стабильное представление, генерируемое FastSpeech 2, обеспечивает надежный вход для ACVAE-VC, позволяя сохранить качество синтеза на уровне, сопоставимом с работой отдельных компонентов, и одновременно повышая общую вычислительную эффективность системы.

В данном разделе был проведён комплексный сравнительный анализ современных алгоритмических подходов в области синтеза речи, реализующих технологии TTS и VC, а также рассмотрены перспективы их интеграции в единую модульную архитектуру. Анализ охватывал ключевые аспекты, такие как точность синтеза, вычислительные затраты, устойчивость к ошибкам, адаптивность и интегрируемость, что позволило выявить существующие сильные и слабые стороны разрозненных методов.

Изучение моделей TTS показало, что современные решения делятся на несколько категорий: End-to-End модели, non-autoregressive модели, Robust TTS решения и Seq2Seq подходы. Каждый из представленных подходов характеризуется своими преимуществами и ограничениями. Так, авторегрессивные модели зачастую обеспечивают высокое качество аудиовыхода, но страдают от увеличенного времени инференса и высокой вычислительной нагрузки, тогда как non-autoregressive решения позволяют существенно ускорить обработку без значительного ухудшения качества.

Аналогичным образом, в области VC рассмотрены различные подходы - от DNN-моделей до архитектур, основанных на генеративно-состязательных сетях и методах обучения на непарных данных. Сравнительный анализ показал, что каждая из этих категорий имеет свои особенности: модели, реализованные на основе глубоких нейронных сетей, способны точно воспроизводить голосовые характеристики, но требуют значительного объёма вычислительных ресурсов, в то время как архитектуры с параллельной генерацией аудиовыхода обеспечивают

более высокую вычислительную эффективность, что особенно важно для систем реального времени. Особое внимание в главе уделено перспективам интеграции решений TTS и VC с использованием единого промежуточного представления и адаптивных методов. Такой подход позволяет не только минимизировать потери информации между этапами обработки, но и обеспечить согласованную работу всех компонентов системы, компенсируя недостатки отдельных моделей. В частности, объединение FastSpeech 2 и ACVAE-VC создаёт синергетический эффект, позволяющий достичь стабильного качества синтеза наряду с улучшенной вычислительной эффективностью.

Интеграция разрозненных методов синтеза речи в единую гибридную архитектуру является перспективным направлением, способным обеспечить высокое качество, устойчивость и адаптивность аудиовыхода. Полученные результаты служат прочной теоретической основой для дальнейшей разработки интегрированной системы синтеза речи, удовлетворяющей требованиям работы в условиях реального времени.

3 Проектирование конвейера синтеза речи

В третьей главе предлагается авторское решение задачи построения модульной системы синтеза речи, состоящей из пяти компонентов: предобработки входных данных, извлечения и формализации признаков, генерации речи, постобработки и адаптации алгоритмов. Задачей практической части является детальное описание структуры системы, функций каждого модуля и порядка передачи данных между ними, а также разработка методических рекомендаций по развертыванию и сопровождению прототипа.

В практической части исследования ставятся две взаимосвязанные задачи: во-первых, спроектировать детальную модульную архитектуру интегрированной системы синтеза речи, объединяющую Text-to-Speech и Voice Conversion, с учётом автоматической адаптации параметров; во-вторых, провести аналитическую апробацию предложенной модели на основе расчётов задержек и оценки ресурсных затрат, опираясь на результаты сравнительного анализа методов из 2 главы. Для достижения этих целей требуется разработать BPMN-диаграмму, описать функциональные задачи каждого из пяти модулей и определить механизмы их взаимодействия и синхронизации без проведения эмпирических экспериментов.

3.1 Проектирование диаграммы гибридной модели

Поскольку выбран гибридный подход к интеграции технологий Text-to-Speech и Voice Conversion, модель синтеза речи должна формироваться так, чтобы критически важные этапы обрабатывались строго последовательно, обеспечивая сохранение акустического и просодического представления, а вспомогательные процессы выполнялись параллельно для оптимизации производительности. Это означает, что этап извлечения и формализации признаков должен предшествовать генерации мел-спектрограммы, а преобразование тембра целевого голоса, в свою

очередь, должно опираться на результаты именно этой формализации. Параллельно со сквозным потоком данных необходимо организовать предобработку входных материалов и динамическую адаптацию параметров, что позволяет поддерживать стабильное качество синтезируемой речи без остановки основных операций.

Важным условием проектирования является единое промежуточное представление в виде мел-спектрограммы: именно она служит связующим звеном между FastSpeech 2 и ACVAEVC, минимизируя потери информации при передаче. Постобработка с внешним вокодером завершается только после того, как мел-спектрограмма целевого голоса сформирована, что гарантирует чистоту и естественность выходного сигнала. Адаптационный модуль, работающий в фоне, аккумулирует логи и метрики, позволяя проводить обновление алгоритмов без влияния на основные задержки.

В предложенной гибридной модели модули разделены на две группы по режиму выполнения. Последовательно обрабатываются три ключевых этапа, которые критичны для сохранения целостности акустического представления:

- Модуль 2 (извлечение и формализация признаков) собирает и упорядочивает duration, pitch и energy;
- Модуль 3 (генерация речи) сначала с помощью FastSpeech 2 создаёт «нейтральную» мел-спектрограмму, а затем ACVAEVC преобразует её в мел-спектрограмму целевого голоса;
- Модуль 4 (постобработка и вокодер) по готовой мел-спектрограмме синтезирует итоговый аудиофайл.

Именно строгая последовательность гарантирует, что каждый набор признаков полностью передаётся следующему этапу без искажений, что крайне важно для передачи тонких просодических и тембровых нюансов.

Одновременно с этим параллельно основному конвейеру работают два вспомогательных модуля:

– Модуль 1 (предобработка входных данных) подготавливает и нормализует новые текстовые или аудио-записи для конвейера, не блокируя работу модулей 2-4;
– Модуль 5 (адаптация и обновление алгоритмов) в фоне аккумулирует логи и метрики всех остальных этапов и, при накоплении достаточного объёма данных, запускает автоматическую подстройку параметров моделей, не замедляя основную обработку.

Такое разделение нагрузок позволяет сочетать высокую точность и надёжность критичных операций с гибкостью и масштабируемостью всей системы и на основании описанной логики был сформирован конвейер синтеза речи, визуализированный в виде BPMN-диаграмм, которые демонстрируют последовательную обработку ключевых этапов и фоновых процедур адаптации.

На рисунке 1 представлен конвейер, получающий на вход текст или аудиофайл, после чего начинается сквозная обработка в пяти логических блоках. Сначала данные нормализуются и приводятся к единому формату: текст разбивается на фонемы, а аудио-на мел-спектрограммы. Затем извлекаются и упорядочиваются ключевые параметры, которые представляются в виде «набора формализованных правил», передающихся на этап генерации речи. На этом этапе FastSpeech 2 сначала создаёт «нейтральную» мел-спектрограмму, а ACVAEVC сразу же превращает её в мел-спектрограмму целевого голоса, сохраняя все интонационные и тембровые особенности оригинала. После этого вокодер преобразует готовую спектрограмму в чистую звуковую волну, и пользователь получает на выходе аудиофайл.

Параллельно с основным конвейером непрерывно работает модуль адаптации: он собирает логи о форматах входных данных, ошибках извлечения признаков, параметрах генерации и показателях качества финального файла, а затем периодически обновляет настройки FastSpeech 2 и ACVAEVC, сохраняя эти параметры в централизованное хранилище.

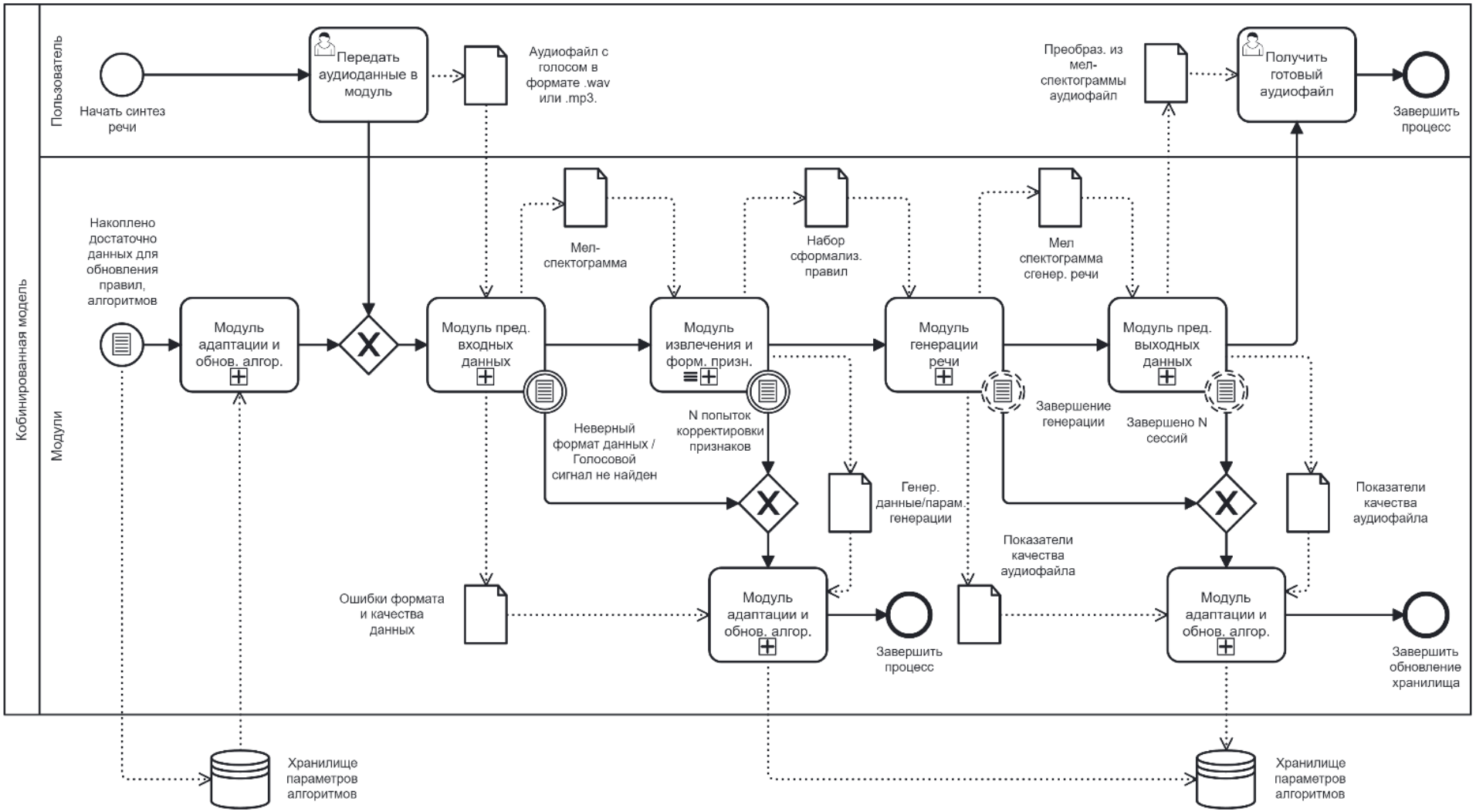

Рисунок 1 - Общий конвейер

Благодаря такому разделению на последовательные задачи конвейера и фоновые процессы, система одновременно гарантирует высокую точность передачи акустического представления и остаётся гибкой к изменениям условий эксплуатации. В итоге получилась архитектура, которая объединяет строгий контроль качества ключевых звеньев с возможностью автоматической подстройки и масштабирования без остановки основного потока обработки.

3.2 Описание модулей и их задач

На рисунке 2 представлен модуль предобработки входных данных, который отвечает за приведение любых пользовательских входных данных к единому формату, необходимому для корректной работы последующих этапов системы. Он принимает на вход либо текстовый файл с исходной фразой, либо аудиозапись в одном из поддерживаемых форматов.

Для текстовых данных выполняется нормализация: удаляются посторонние символы и метатеги, приводится весь текст к нижнему регистру, после чего строка разбивается на фонемы. Это обеспечивает однородную фонемную основу, что критично для точного сопоставления с акустическими моделями. При обработке аудиофайлов модуль проверяет параметры записи - частоту дискретизации, число каналов, уровень шума и громкость - и при необходимости автоматически конвертирует файл в моно, фильтрует фоновый шум и выравнивает громкость по заданному порогу.

Для реализации всех этих операций модуль предобработки входных данных может опираться на любые надёжные библиотеки и инструменты общего назначения, поддерживающие нормализацию текста, фонемизацию, конвертацию аудиоформатов, ресэмплинг и базовую фильтрацию шума. Такой подход позволяет в дальнейшем без дополнительных затрат менять или обновлять используемые средства в зависимости от условий эксплуатации и новых требований.

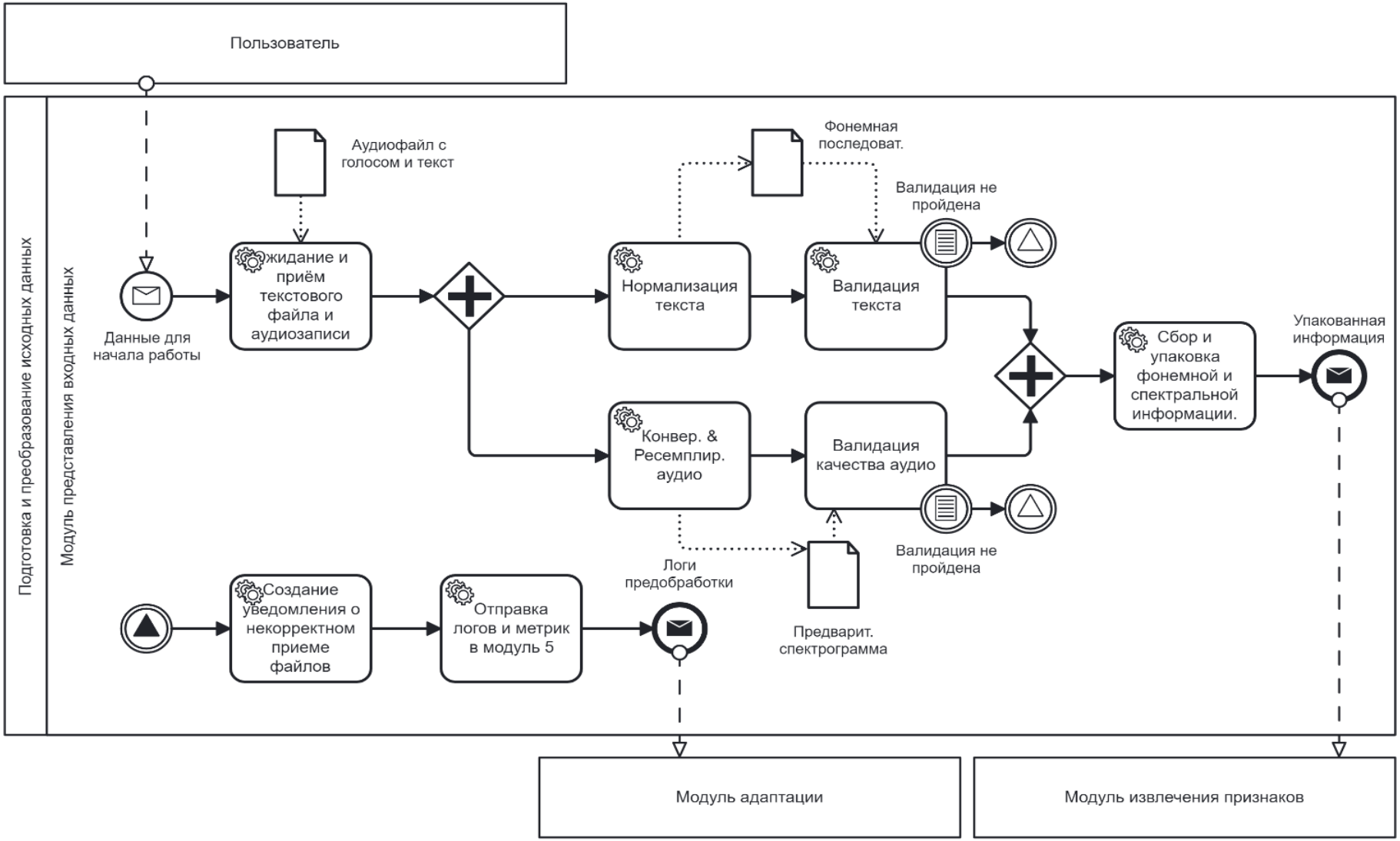

Рисунок 2 - Модуль предобработки входных данных

Модуль 1 спроектирован так, чтобы одновременно поддерживать три режима работы: «только текст», «только аудио» и «гибридный» («текст + запись голоса»). На диаграмме модуля 1 заложен процесс одновременного приёма текстового файла и аудиозаписи, чтобы отобразить гибкость модели. Сразу после старта процесса система проектирует ветвление, когда оба потока запускаются параллельно. Такое решение гарантирует, что при наличии и текста, и примера голоса получают полное представление-фонемную разметку от текстового сегмента и спектральную информацию от аудиопотока-что жизненно важно для точного интегрированного синтеза.

Нормализация текста начинается с удаления лишних символов, таких как управляющие коды, HTML-теги или случайные артефакты копирования. Затем система приводит все символы к нижнему регистру и прогоняет текст через фонемизатор, сопоставляющий каждое слово с его звуковой транскрипцией. Внутри этой задачи может задействоваться словарь фонем или статистическая модель, но главное-чтобы на выходе получилась однородная последовательность фонем, готовая к дальнейшей работе модулей TTS.

В задаче конвертации и ресэмплинга аудио исходная запись в формате WAV или MP3 приводится к моно-режиму с частотой дискретизации 16 кГц, прогоняется через фильтр подавления фонового шума и нормализуется по уровню громкости. Результатом этой цепочки является мел-спектрограмма, матрица, готовая для передачи во второй модуль. После этих предварительных шагов срабатывают точечные валидации: фонемная разметка проверяется на корректность, а мел-спектрограмма - по соотношению «сигнал/шум» и длительности. В случае провала любой из проверок управление передаётся в ветвь обработки ошибок, где создаётся уведомление о некорректном приёме, формируются логи и показатели и отправляются в модуль 5 для дальнейшего анализа и автоматической корректировки порогов. Основной же поток при этом не останавливается и ждёт

новых данных. Когда обе ветви успешно завершаются, их результаты упаковываются в два определённых артефакта:

– Файл с расширением .phn, содержащий список фонем и, при необходимости, временные метки.
– Файл с расширением .melspec, хранящий мел-спектрограмму в виде бинарного массива или стандартного формата.

Эти два артефакта передаются далее как вход в модуль 2, где из .phn формируются численные векторы длительностей, высот и энергий, а из .melspec - спектральные эмбеддинги для генерации речи.

Механизмы валидации качества отслеживают, чтобы после всех преобразований сигнал оставался чистым, а текст не содержал недопустимых символов. В случаях, когда сигнал слишком зашумлён или в тексте обнаружены ошибки, модуль формирует служебное сообщение и передаёт его в модуль 5 для дальнейшего анализа и накопления статистики.

Благодаря подобной организации, модуль 1 одновременно остаётся простым и надёжным: он выполняет все задачи по нормализации и валидации, аккуратно отсекая «плохие» входы и передавая «хорошие» данные в строго определённом формате. Последовательная подготовка и параллельное логирование-обеспечивает непрерывность основного конвейера и даёт возможность фоновой адаптации без остановки процесса синтеза речи.

Модуль 2 отвечает за извлечение и формализацию признаков из подготовленных фонемных и спектральных данных, полученных на выходе Модуля 1. Его основная задача - преобразовать «сырые» последовательности фонем и мел-спектрограмм в численные векторы параметров, пригодных для последующей генерации речи и конверсии голоса. В рамках этого процесса каждый сегмент фонемы сопоставляется с набором временных меток и длительностей, а каждая мел-спектрограмма анализируется на предмет ключевых акустических

характеристик: частотной огибающей, тона (F0), энергии и спектральной плотности.

На рисунке 3 представлена работа модуля 2, которая строится по принципу двухэтапного конвейера: сначала выполняется детальное извлечение низкоуровневых признаков - длительности фонем, контуров F0 и кривой энергий по кадрам, затем эти признаки агрегируются и нормализуются в единую многомерную матрицу или набор тензоров, готовых для входа в TTS- и VC-модели. Такой подход позволяет свести разнородные данные к унифицированному формату и одновременно оставить возможность гибко настраивать веса отдельных параметров при обучении или инференсе.

Для реализации Модуля 2 широко применяются проверенные библиотеки цифровой обработки сигналов и анализа речи. Так, извлечение контуров F0 и энерго-информации может выполняться средствами pyworld или Parselmouth, а спектральные признаки - через librosa или Kaldi-инструменты. Формализация и упаковка данных в тензоры обычно выполняется с помощью NumPy и TensorFlow/PyTorch.

Важность Модуля 2 состоит в том, что именно он формирует «мост» между необработанным аудиосигналом и высокоуровневыми нейросетевыми архитектурами TTS и VC. От качества извлечённых и нормализованных признаков напрямую зависит точность интонации, естественность тембра и мобильность адаптации голоса к новым условиям.

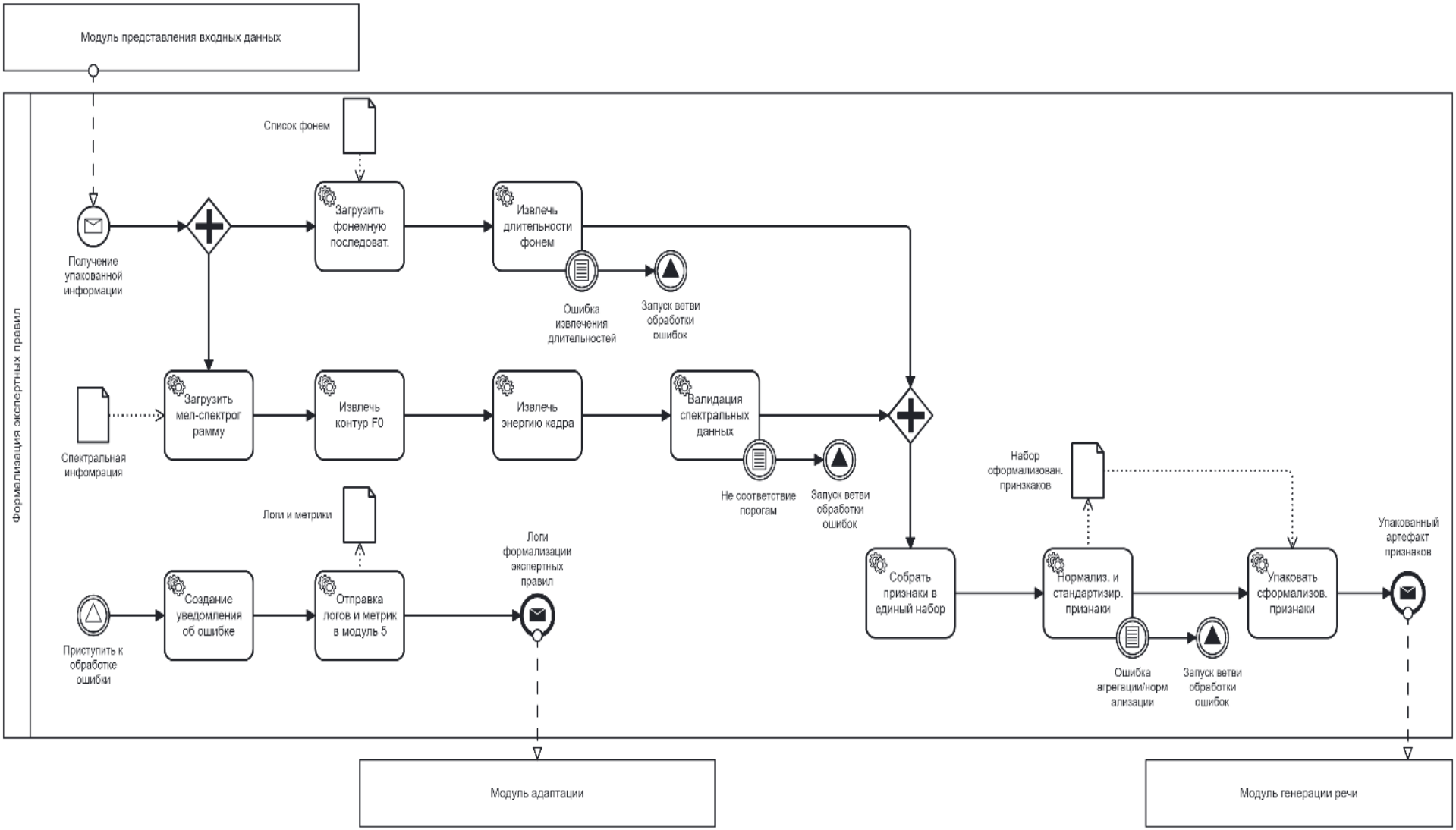

Рисунок 3 - Модуль извлечения и формализации признаков

В результате работы Модуля 1 на выходе формируются два основных артефакта - файл с фонемной разметкой (.phn), в котором каждая фонема снабжена временной меткой, и файл мел-спектрограммы (.melspec), отражающий спектральное содержание исходного аудио. Эти артефакты поступают на вход Модуля 2 «Извлечение и формализация признаков», где сразу же разбиваются на две параллельные ветви обработки: фонемную и спектральную.

В фонемной ветви первым шагом служит загрузка фонемной последовательности из .phn. На её основе выполняется вычисление длительности каждой фонемы - разница между начальной и конечной меткой дает значение в миллисекундах. Если в файле отсутствуют временные метки или метки лежат за пределами допустимого диапазона, система фиксирует ошибку извлечения длительностей: генерируется подробный лог с указанием невалидных записей и запускается саб-процесс уведомления об ошибке, в котором сообщение обрабатывается Модулем 5. Это позволяет избежать каскадных сбоев на более поздних этапах и собрать статистику по частым артефактам фонемной разметки.

Одновременно во второй, спектральной ветви, происходит загрузка мел-спектрограммы из файла .melspec. Первой задачей здесь является извлечение контура основного тона (F0) - алгоритмы PyWorld или Praat оценивают периодичность огибающей сигнала по кадрам, что важно для передачи интонации и эмоциональной окраски голоса. Далее вычисляется энергия каждого звукового кадра - суммарная амплитуда всех частот в заданном окне. При обнаружении некорректных значений срабатывает валидация спектральных данных: ошибки фиксируются в логах, и аналогично фонемной ветви запускается ветвь уведомления об ошибках.

После того как обе ветви успешно завершили свои операции, их результаты объединяются через объединительный шлюз. В этот момент запускается сервис-таск «Сбор признаков в единый набор», который формирует структуру данных, включающую векторы длительностей, значения F0 и энергии кадра.

Полученный «набор формализованных признаков» затем передаётся на этап нормализации: все три вектора приводятся к единому масштабу, благодаря чему downstream-модули получают согласованные по распределению и диапазону входные данные.

Если на этапе агрегации или нормализации возникает ошибка, срабатывает соответствующая проверка, и процесс уходит в ветвь обработки ошибок - генерируется лог агрегации/нормализации, после чего логи отправляются в Модуль 5 для анализа и дальнейшей настройки параметров системы.

«В качестве выходного артефакта Модуля 2 используется HDF5-файл <basename>.h5, внутри которого расположены пять ключевых датасетов и набор атрибутов метаданных:

- /mel_spectrogram (тип float32, размерность [T × 80]) - мел-спектрограмма (FFT=1024, hop_length=256, win_length=1024).
- /duration (тип int32, размерность [N]) - длительности каждой из N фонем в кадрах.
- /pitch (тип float32, размерность [T]) - контур основного тона (F0) по кадрам.
- /energy (тип float32, размерность [T]) - энергия каждого аудиокадра.
- /phonemes (тип строка, размерность [N]) - список фонем, соответствующий датасету /duration.

Атрибуты корня файла (sampling_rate=16000, hop_length=256, win_length=1024, num_mel_bins=80) гарантируют самодостаточность формата и позволяют downstream-модулям (FastSpeech 2, ACVAEVC) однозначно интерпретировать данные.» Готовый файл передаётся в Модуль 3 посредством конечного сообщения, обеспечивая непрерывный конвейер обработки: downstream-модуль получает полный, согласованный набор просодических и спектральных признаков для дальнейшей генерации или конверсии речи.

Модуль 2 предоставляет надёжное и детализированное извлечение всех ключевых акустических и просодических характеристик речи, строгий контроль качества на каждом шаге и формирование стандартизованного входного пакета для следующих этапов синтеза.

Модуль 3 «Генерация речи» является центральным в конвейере интегрированного синтеза, поскольку именно здесь объединяются технологии Text-to-Speech (TTS) и Voice Conversion (VC) для получения окончательной мел-спектрограммы целевого голоса. На его вход поступает стандартизованный HDF5-файл из Модуля 2, содержащий мел-спектрограмму, фонемные длительности, контур тона и энергию; на выходе формируется мел-спектрограмма, адаптированная под индивидуальные особенности говорящего.

На рисунке 4 представлена работа Модуля 3, которая строится по принципу двух последовательно связанных этапов. Сначала TTS-компонент на основании фонемных длительностей и исходной мел-спектрограммы восстанавливает высококачественный «нейтральный» спектрограммовый сигнал, выровненный по тексту и временным меткам. Затем VC-блок, получив этот «нейтральный» спектрограммовый вход, преобразует его в спектрограмму, сохраняющую тембровые и просодические особенности целевого говорящего. Такой подход позволяет обеспечить точное текст-ориентированное выравнивание и при этом привнести в итоговую речь характерные интонационные и тембральные нюансы без необходимости иметь большой корпус голосовых записей одного и того же говорящего.

Для реализации Модуля 3 полходят две экосистемы - PyTorch и TensorFlow 2. PyTorch обеспечивает гибкую исследовательскую среду с динамическими вычислительными графами и множеством community-реализаций FastSpeech 2, а TensorFlow 2 в сочетании с TensorFlowTTS предоставляет готовые к промышленному развертыванию модели и инструменты оптимизации для мобильных и серверных платформ.

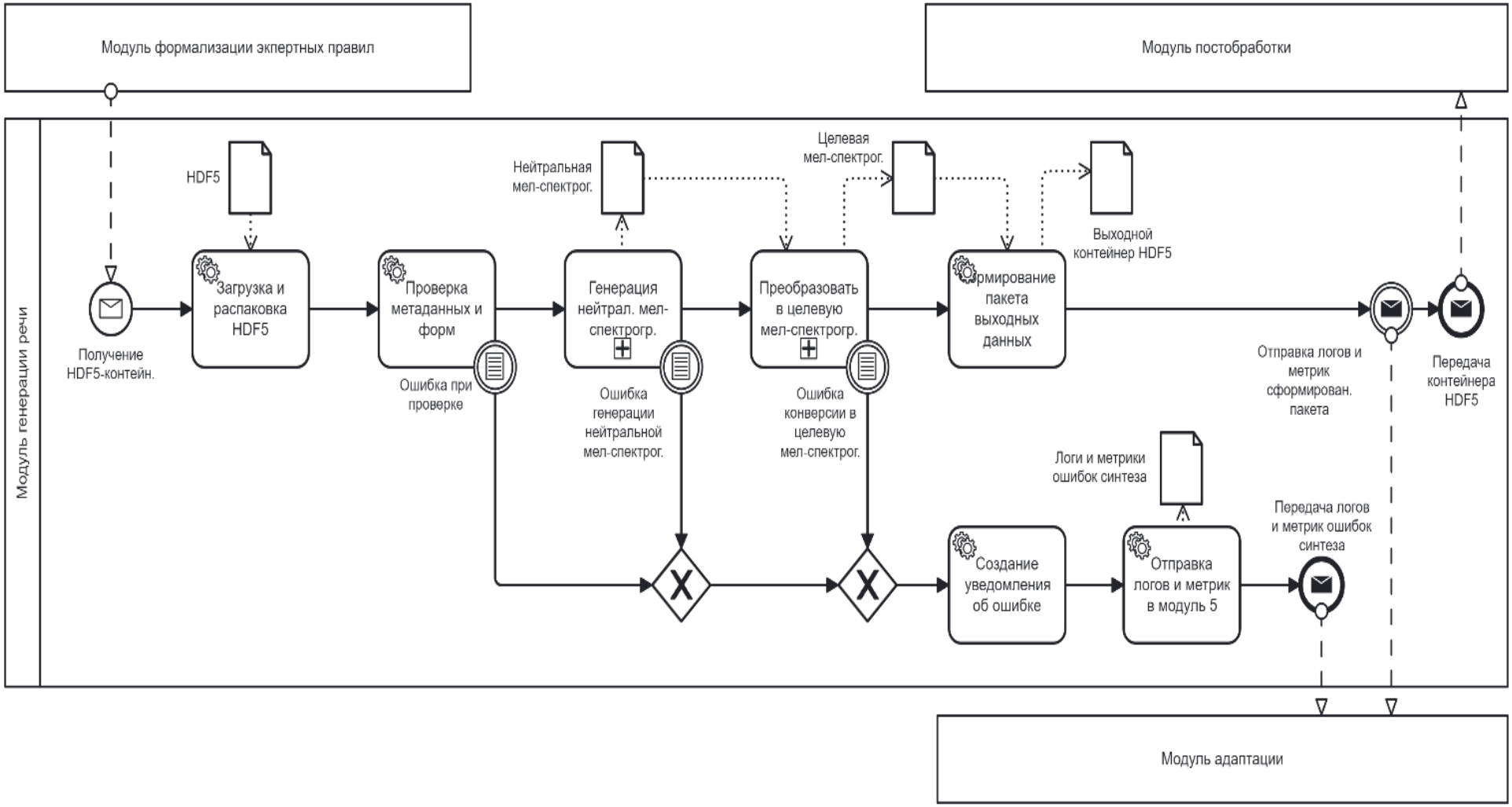

Рисунок 4 - Модуль генерации речи

В Модуле 3 «Генерация речи» поступивший из Модуля 2 HDF5-контейнер выступает единственным источником всех необходимых данных: в нём хранятся оригинальная мел-спектрограмма, длительности фонем, контур высоты тона и энерго-контур. Первым шагом внутри Модуля 3 выполняется верификация структуры контейнера: проверяются наличия обязательных датасетов и согласованность их размеров, после чего формируется объект Input_HDF5, готовый к подаче на этап TTS.

Далее с помощью FastSpeech 2 происходит восстановление «нейтральной» мел-спектрограммы. На этом этапе модель читает фонемные длительности и исходный спектрограммовый сигнал, выравнивает его по текстовым меткам, прогнозирует недостающие кадры и составляет гладкую спектрограмму, лишённую говоровых особенностей исходного диктора. Результатом работы FastSpeech 2 становится новый dataset /neutral_mel внутри контейнера - Neutral_HDF5 - который отличается от входного добавленным массивом мел-коэффициентов с чётко выровненной по времени формой.

Следующим этапом в том же контейнере запускается блок Voice Conversion, реализованный на основе ACVAEVC. Он принимает /neutral_mel и, используя ранее извлечённые просодические векторы (/pitch, /energy), адаптирует спектрограмму так, чтобы она приобрела тембровые и интонационные особенности целевого говорящего. Внутри контейнера создаётся новый датасет /target_mel, и на выходе модуля формируется Output_HDF5, содержащий уже три спектрограммы: исходную, нейтральную и целевую.

В случае возникновения ошибок на любом из этапов, записываются коды ошибок, имя не прошедшего этапа, timestamp и URI контейнера. Логи и метрики ошибок синтеза отправляются в Модуль 5, где накапливаются статистика отказов и автоматически корректируются параметры моделей для последующих запусков.

При успешном завершении конверсии запускается финальный процесс «Формирование пакета выходных данных». Он упаковывает весь контейнер с новыми датасетами в единый архивный файл Final_HDF5 и передаёт его в модуль 4. Одновременно отправляются логи и метрики синтеза с такими метриками, как latency каждого шага, размеры спектрограмм и качество предварительной выборки в модуль адаптации, что обеспечивает полноценную обратную связь для адаптации и будущего улучшения конвейера.

На рисунке 5 представлен модуль постобработки, который отвечает за преобразование итоговой мел-спектрограммы целевого голоса в аудиосигнал, готовый к воспроизведению. Несмотря на то, что в модуле 3 происходит объединение Text-to-Speech и Voice Conversion для генерации «нейтральной» и адаптированной мел-спектрограммы, непосредственное превращение спектрограммы в звуковую волну выносят отдельно. Такой подход обеспечивает разделение ответственности: модуль 3 сосредоточен на формировании акустических признаков, а модуль 4 - на генерации качественного аудиовыхода с учётом требований к задержкам и аппаратным ресурсам .

Выделение вокодирования в отдельный модуль обусловлено рядом технических и архитектурных факторов. Во-первых, современные нейронные вокодеры требуют собственных механизмов оптимизации, которые отличаются от механизмов обучения TTS и VC. Второй этап модели часто опирается на GAN-структуры или потоковые модели, что предъявляет особые требования к GPU-памяти и скорости инференса. Во-вторых, отделение вокодера позволяет независимо обновлять или заменять его реализацию без необходимости переобучать модуль генерации мел-спектрограмм. На практике оптимальным выбором часто становится HiFi-GAN благодаря сбалансированному сочетанию качества и скорости обработки.

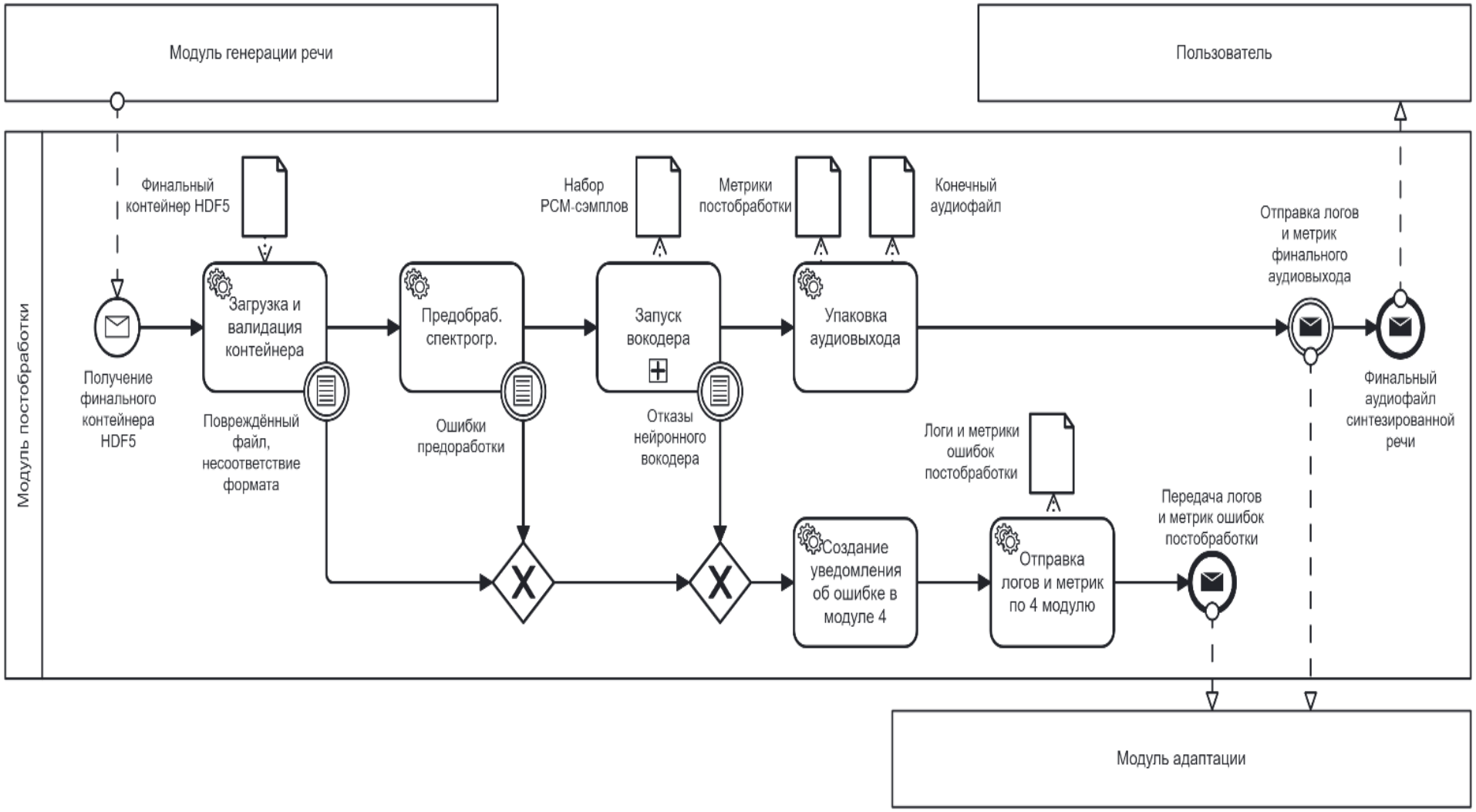

Рисунок 5 - Модуль постобработки

В качестве входного артефакта Модуля 4 используется окончательный HDF5-контейнер, полученный из Модуля 3, - файл с расширением .h5, в корне которого должны присутствовать следующие датасеты и атрибуты метаданных, обеспечивающие полную самодостаточность:

- /mel_spectrogram_target (тип float32, размерность [T×80]) - целевая мел-спектрограмма с учётом тембровых и просодических особенностей говорящего, полученная после VC-этапа.
- /mel_spectrogram_neutral (тип float32, размерность [T×80]) - «нейтральная» спектрограмма, сгенерированная FastSpeech 2.
- /duration (тип int32, размерность [N]) - длительности фонем в кадрах, синхронизированные с обеими спектрограммами.
- /pitch (тип float32, размерность [T]) и /energy (тип float32, размерность [T]) - векторы F0 и энергий по каждому кадру.
- /phonemes (тип строка, размерность [N]) - список фонем, соответствующий датасету /duration.

Атрибуты корневой группы (sampling_rate, hop_length, win_length, num_mel_bins) позволяют downstream-модулю однозначно трактовать данные без внешних справочных файлов.

Загрузка и валидация контейнера.

При получении HDF5-файла из Модуля 3 запускается процедура целостной проверки структуры: вначале файл монтируется в виртуальной файловой системе, затем с помощью HDF5-API проверяется наличие всех обязательных групп и соответствие их размерностей заявленным в атрибутах. Если обнаруживается отсутствие датасета или несоответствие размерности, то срабатывает шлюз обработки ошибок. В этом случае в задаче «Создание уведомления об ошибке» формируется подробный лог, содержащий имена отсутствующих или неправильно размерных датасетов, и отправляется в Модуль 5 для дальнейшего анализа.

Основной поток при этом переводится в состояние «Ошибка валидации» и прерывается, чтобы избежать ложного синтеза из повреждённых данных.

После успешного прохождения валидации контейнер распаковывается в оперативную память, и над датасетом /mel_spectrogram_target выполняется серию фильтрационных и нормализующих операций. В частности, для устранения «звонких» артефактов могут применяться полосовые фильтры на частотах вне человеческой речи, а динамический диапазон каждого кадра выравнивается по алгоритму автоматической регулировки усиления (AGC). Кроме того, для повышения устойчивости к шумам выполняется мягкое сглаживание соседних мел-коэффициентов через скользящее окно. Если на этом этапе обнаруживается чрезмерное отклонение амплитудных значений за допустимые пороги, запускается ветвь обработки ошибок, аналогично предыдущей, и формируется лог «Метрики постобработки», отправляемый в Модуль 5.

С очищенной и нормализованной мел-спектрограммой модуль переходит к синтезу звуковой волны. Здесь инкапсулирован в сложный саб-процесс нейронный вокодер - чаще всего это HiFi-GAN из-за его способности сохранять высокую частотную детализацию, либо Parallel WaveGAN для ускорённого инференса. Суб-процесс принимает на вход массив размерности [T×80] и внутренние атрибуты контейнера (частота дискретизации, hop_length), обрабатывает каждый кадр с учётом временных контекстов и выдаёт одномерный массив PCM-сэмплов (int16 или float32). При ошибках во время генерации вокодер возвращает сигнал исключения, который улавливается шлюзом «Ошибка вокодирования», и далее создаётся уведомление и лог «Логи ошибок постобработки» для Модуля 5.

Полученный из саб-процесса массив PCM-сэмплов далее оборачивается в стандартный WAV-контейнер. В задаче «Упаковка аудиовыхода» к потоку байтов добавляется RIFF-заголовок с указанием параметров (FmtChunk, DataChunk), записываются метаданные (дата, идентификатор процесса) и рассчитываются отчётные метрики выходного файла: пиковое значение, среднеквадратичный

уровень, спектральная плотность мощности. Итоговая структура представляется как файл output.wav, а собранные метрики сериализуются в JSON-объект «Метрики постобработки» и передаются в модуль адаптации

В случае возникновения ошибок на любых этапах постобработки Модуля 4 формируется специализированный пакет логов и метрик, который передаётся в Модуль 5 для централизованного анализа и адаптации. Ниже приведены уточнённые механизмы:

а) Если при проверке структуры HDF5-контейнера обнаруживается отсутствие или неправильная размерность датасета, в пакет «Логи валидации HDF5» включаются:
   1) Наименование и путь к отсутствующему или некорректному датасету;
   2) Ожидаемая и фактическая размерности;
   3) Контрольная сумма корневой группы (CRC32) для быстрой диагностики целостности файла;
   4) Время и идентификатор задачи, в которой произошла ошибка. Этот пакет передаётся в Модуль 5 через событие Message Error «HDF5 Validation Failed».

б) При выявлении чрезмерного уровня шума или аномальных отклонений в амплитудных значениях в задаче предобработки генерируется пакет «Метрики постобработки», содержащий:
   1) Среднеквадратичный уровень (RMS) по кадрам и его отклонение от опорного RMS;
   2) Статистику пиковых значений (max, min) и процент кадров, превышающих допустимый порог (±6 дБ);
   3) Список применённых фильтров и их параметры;
   4) Фрагмент спектрограммы для визуальной инспекции. Одновременно запускается ветвь «Ошибка постобработки» через

Exclusive Gateway, и метрики передаются в Модуль 5 по событию Message Error «Spectrogram Preprocessing Failed».

в) Если нейронный вокодер (HiFi-GAN или Parallel WaveGAN) выдаёт «мёртвую» волну или заметные цифровые артефакты, формируется пакет «Логи ошибок вокодера»:
   1) Параметры модели и checkpoint (хеш совпадающего файла весов);
   2) Средняя энтропия выходного сигнала (мерило гладкости);
   3) Спектральная плотность мощности на контрольных частотах (0-8 кГц);
   4) Фрагмент PCM-данных (несколько тысяч сэмплов) для отладки. Это приводит к активации шлюза «Ошибка вокодирования» и передаче данных в Модуль 5 посредством Message Error «Vocoder Failure».

г) При корректном генерировании WAV-файла параллельно ему формируется пакет «Метрики постобработки», включающий:
   1) Общий RMS и пиковый уровень выходного файла;
   2) Длительность в секундах и соответствие заявленной частоте дискретизации;
   3) Состояние целостности RIFF-заголовка (CRC);
   4) Статистику спектральной плотности мощности на ключевых диапазонах речи (300-3400 Гц, 3400-8000 Гц). Пакет отправляется в Модуль 5 через событие Message «Postprocessing Metrics».

Таким образом, каждая ветвь ошибок снабжается собственным набором логов и метрик, что позволяет Модулю 5 не только фиксировать сбои, но и автоматически корректировать параметры в последующих запусках, избегая повторения тех же исключений и повышая общую надёжность конвейера.

Когда Модуль 4 завершает конвейер интегрированного синтеза: он не только превращает согласованные спектрограммы в воспроизводимый аудиосигнал, но и обеспечивает многоуровневую валидацию, предобработку и логирование, создавая

надёжный фреймворк для дальнейшего мониторинга и автоматической адаптации всей системы.

В рамках гибридной архитектуры Модуля «Адаптация и обновление алгоритмов» ключевым становится не только сама процедура дообучения нейронных сетей, но и тщательное управление объёмом и качеством накопленных диагностических данных. Этот модуль выполняет функцию центрального аналитического узла: все журналы ошибок и измеренные метрики, порождаемые на этапах предобработки, извлечения признаков, генерации мел-спектрограмм и вокодирования, поступают сюда и аккумулируются в едином централизованном хранилище. Такое решение позволяет не просто хранить сухие логи, а классифицировать их по ключевым направлениям - валидация входных контейнеров, сбои генерации «нейтральных» спектрограмм, ошибки конверсии целевой мел-спектрограммы, отказ нейронного вокодера и т. д. - и на их основе выстраивать динамические правила адаптации.

Трансферное обучение в задачах TTS и VC традиционно предполагает тонкую настройку предобученных моделей на большом количестве размеченных записей целевого говорящего. Исследования показали, что «для достижения качества, сопоставимого с обучением «с нуля» на масштабных корпусах, требуется не менее 30 минут чистых аудиозаписей одного говорящего. При сокращении объёма адаптационных данных до 10 минут наблюдается значимое падение субъективных и объективных метрик (MOS, PESQ, STOI)» [24]. Хотя «использование адаптеров позволяет снизить порог до примерно 5 минут без заметных потерь качества» [10], такие методы требуют модификации архитектуры и усложняют процесс масштабирования.

В связи с тем, что наша система рассчитана на быстрый запуск и работу с минимальными объёмами записей (1-2 минуты речи), при построении модели было решено отказаться от традиционного трансферного обучения и сосредоточиться на методах self-supervised learning, способных извлекать устойчивые фонемно-

спектральные представления из любых неразмеченных аудиоданных. Такое решение позволяет обеспечить высокую естественность синтеза даже при минимальных требованиях к размеченным материалам и значительно упростить организацию сбора данных.

Методы self-supervised learning открывают возможность непрерывного извлечения фонемно-спектральных представлений из любых аудиоданных без необходимости ручной разметки. При использовании SSL-моделей система обучается на «сырых» записях, автоматически выстраивая устойчивые к шумам и акцентным вариациям эмбеддинги, которые затем интегрируются в конвейер FastSpeech 2 → ACVAEVC. Это позволяет обеспечить высокую естественность синтезированной речи даже при минимальном объёме адаптационных примеров и в условиях разнообразной акустической обстановки.

На рисунке 6 представлена работа Модуля 5, которая организована как непрерывный цикл «сбор - анализ - адаптация - деплой», в рамках которого аккумулируются все логи ошибок и объективные метрики работы конвейера, после чего на их основе запускаются целенаправленные процедуры дообучения self-supervised модели.

Сразу после каждого полного прохода синтеза в Модуль 5 поступают три группы артефактов. Первая группа - пакет логов ошибок, где для каждого сбоя фиксируются исходный модуль (1-4), категория ошибки и служебная диагностическая информация. Вторая группа - набор объективных метрик финального аудиофайла, сформированных в Модуле 4: длительность WAV и проверка частоты дискретизации, RMS и пиковый уровень, SNR, спектральная плотность мощности в ключевых диапазонах речи, целостность RIFF-заголовка, объём и структура FmtChunk/DataChunk. Третья группа - метрики успешных операций, охватывающие проверку структуры HDF5-контейнера, статистику извлечённых признаков и валидацию «нейтральной» и «целевой» спектрограмм на предмет равномерности амплитуды и отсутствия пропусков.

Все эти данные централизуются в хранилище и автоматически классифицируются по направлениям: валидация входных данных, извлечение и агрегация признаков, генерация спектрограмм TTS/VC и постобработка с вокодированием. Такое детальное разбиение позволяет точно локализовать «узкие места» конвейера и формировать «триггерные условия» для запуска дообучения: при росте числа ошибок извлечения F0 инициируется переобучение SSL-модели на фрагментах с богатой интонацией, при снижении SNR - фокус на шумовые примеры.

Для каждого типа артефакта в хранилище ведётся отдельный счётчик накопленных «чистых» примеров. Например, для фонемно-спектральных эмбеддингов выставлен порог 30 000 сегментов по 1 секунде, эквивалент 30 одно-секундных фрагментов разных голосов, а для шумовых фрагментов - 10 000 сегментов в диапазоне SNR < 20 дБ. Как только любой из счётчиков достигает своего порога, генерируется системное событие «Data Threshold Reached» с указанием категории данных и точным объёмом накопленного материала. Это событие поступает в Оркестратор обучения.

Оркестратор извлекает из хранилища только те записи, которые относятся к сработавшему триггеру, вместе с сопутствующими метриками и контекстом ошибок. К примеру, при достижении порога фонемно-спектральных эмбеддингов выбираются те фрагменты, где были зарегистрированы систематические отклонения контура F0; при пороге шумовых сегментов - фрагменты с SNR < 20 дБ. Затем на основании характера накопленных данных автоматически формируются два типа задач:

В результате Модуль 5 обеспечивает непрерывное улучшение интегрированного конвейера синтеза речи: система автоматически реагирует на любые отклонения в качестве выходного аудио, не требует длительных сессий записи целевого говорящего и сохраняет гибкость к изменению акустических условий.

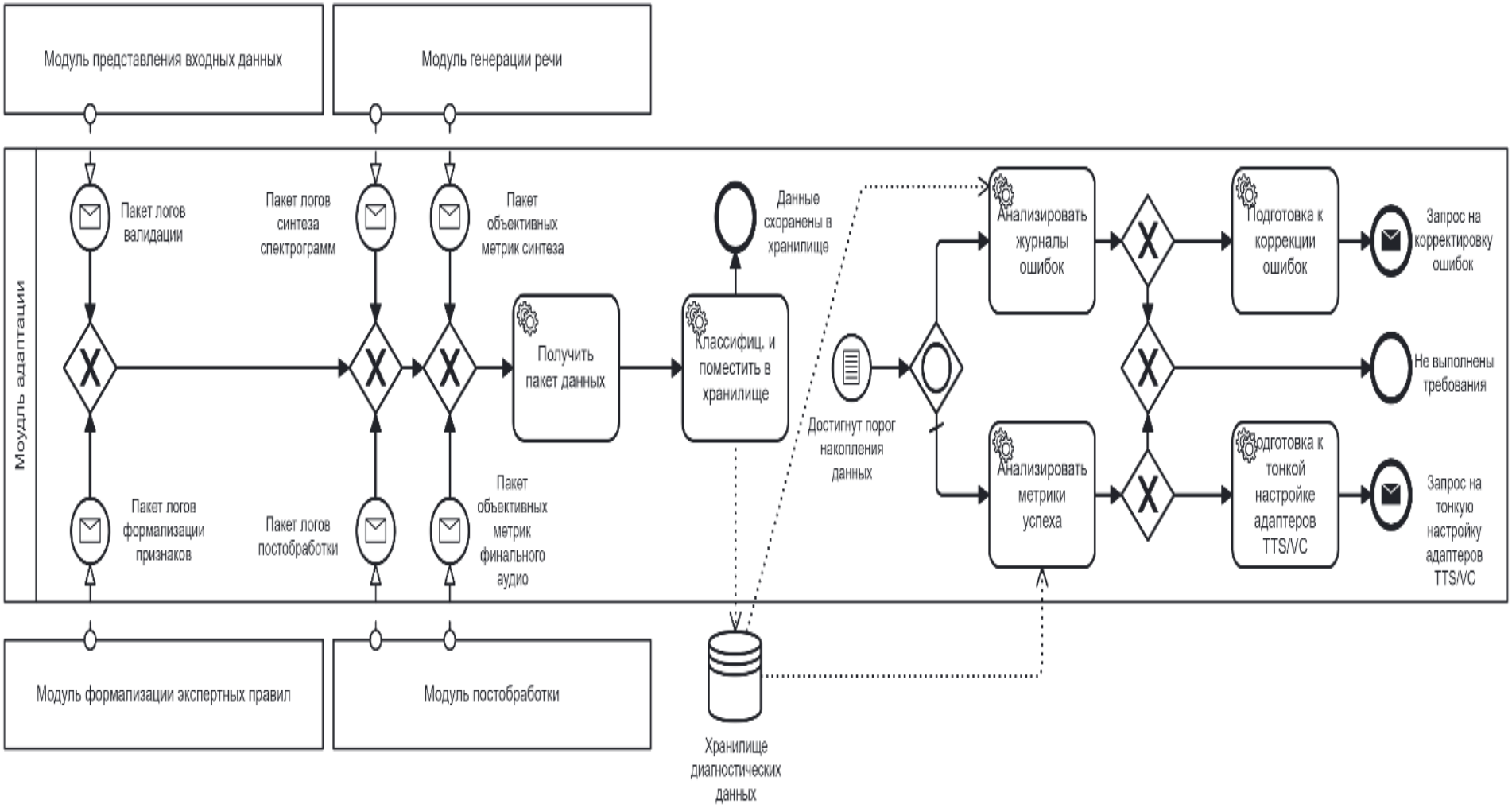

Рисунок 6 - Модуль адаптации и обновления алгоритмов

Первым этапом работы Модуля 5 является приём информации, поступающей от остальных четырёх блоков конвейера синтеза речи, и её первичная унификация в виде шести типов входных пакетов. Такое разделение позволяет сохранить полную трассировку происходящих в системе событий, гарантировать целостность диагностических данных и подготовить их для последующего анализа и хранения.

Во-первых, каждый из четырёх рабочих модулей (предобработка, извлечение признаков, генерация спектрограмм и постобработка) по завершении своей работы формирует пакет логов ошибок в случае любого сбоя или некорректного исходного состояния. Во-вторых, из Модуля 4 дополнительно поступают два «успешных» пакета: набор объективных метрик по финальному аудиовыходу и пакет контролируемых метрик промежуточных этапов (валидность HDF5-контейнера, качество спектрограмм).

Входящие в Модуль 5 данные подразделяются на шесть категорий:

- Логи ошибок Модуля 1 содержат сведения о некорректной нормализации текстовой информации или неверных параметрах аудиозаписи (частота дискретизации, моно/стерео, уровни шума, нераспознанные символы), а также диагностическую информацию об источнике сбоя.
- Логи ошибок Модуля 2 фиксируют сбои при извлечении фонемных длительностей, при агрегации и нормализации спектральных признаков и при нарушении пороговых условий.
- Логи ошибок Модуля 3 описывают неудачи на этапах генерации «нейтральной» мел-спектрограммы FastSpeech 2 или преобразования её в целевую форму ACVAEVC, а также ошибки валидации метаданных входного HDF5-контейнера.
- Логи ошибок Модуля 4 включают информацию о неверной структуре или повреждении входного HDF5-файла, сбоях в предобработке спектрограммы и отказах нейронного вокодера при синтезе WAV-сигнала.

– Объективные метрики финального аудио собираются в процессе упаковки выходного аудиофайла - в них входят длительность, проверка частоты дискретизации, RMS-уровень, пиковые значения, SNR, спектральная плотность мощности в ключевых диапазонах, целостность RIFF-заголовка и структура FmtChunk/DataChunk.
– Метрики промежуточных успешных операций представляют собой контролируемые показатели корректности HDF5-контейнера, соответствие размерностей и атрибутов, а также статистику извлечённых признаков и валидацию равномерности амплитудной огибающей в «нейтральной» и «целевой» спектрограммах.

Каждый из перечисленных пакетов передаётся в Модуль 5 по механизму messaging, где первым шагом выполняется их приведение к единому внутреннему формату. Независимо от источника, все пакеты снабжаются обязательными метаданными:

– source_module - идентификатор исходного модуля (1-4),
– message_type - категория пакета (ошибка или метрики),
– timestamp - метка времени окончания операции,
– session_id - сквозной идентификатор конвейерного прохода,
– payload - собственно содержимое лога или метрик.

Текстовые логи и диагностические сообщения нормализуются: приводятся к единой кодировке UTF-8, избыточные символы удаляются, стек вызовов при необходимости обрезается до предопределённой глубины, а имена файлов и путей переписываются в относительный вид. Метрики в виде числовых значений упаковываются в заранее описанные JSON-объекты со строгими схемами валидации, что позволяет гарантировать однозначность их интерпретации downstream-процессами.

После нормализации каждый из шести типов пакетов (логи ошибок и два вида метрик от Модулей 1-4) поступает в секцию классификации Модуля 5, где сразу же присваиваются следующие атрибуты:

- category - направление,
- status - «ошибка» или «успех»,
- source_module - номер модуля-источника,
- timestamp и session_id - для сквозного трекинга.

На основании этих атрибутов пакеты либо отправляются в хранилище ошибок (для дальнейшего анализа срабатываний), либо в хранилище метрик успеха (для мониторинга качества).

Каждый пакет сразу же поступает в централизованное хранилище данных, реализованное в виде набора специализированных таблиц (или коллекций), где:

- error_logs хранит все сведения об ошибках (source_module, category, timestamp, session_id, diagnostic_info),
- audio_metrics аккумулирует объективные параметры финального WAV-файла (duration, sampling_rate, RMS, peak_level, SNR, spectral_density, RIFF_integrity),
- intermediate_metrics содержит результаты проверки промежуточных этапов (HDF5_validation, feature_statistics, spectrogram_quality).

Каждая запись в этих таблицах включает поля category, status, source_module, timestamp и session_id, что обеспечивает точную фильтрацию и выборку данных для последующего анализа и запуска адаптационных процедур.

Централизованное хранилище Модуля 5 выполняет функцию надёжного репозитория для всех логов и метрик, поступающих из этапов предобработки, извлечения признаков, генерации мел-спектрограмм и вокодирования. Архитектурно оно спроектировано на основе сочетания документного хранилища для неструктурированных логов и реляционной базы данных для строго типизированных метрик. В документном слое сохраняются подробные трассировки

ошибок, содержащие параметры входных контейнеров, стек вызовов и контекст задачи. В табличном слое фиксируются числовые значения показателей качества выходных аудиофайлов, таких как RMS-уровень, SNR и характеристики заголовков WAV-контейнеров. Отдельный набор таблиц или BLOB-полей предназначен для хранения SSL-эмбеддингов - массивов фонемно-спектральных представлений фиксированной размерности. Корневая схема метаданных обеспечивает указание версии моделей, контрольных сумм и параметров конвейера, что гарантирует согласованность интерпретации данных в последующих процессах анализа и адаптации.

Входящие пакеты логов и метрик проходят через многоступенчатую логику классификации, в результате чего каждый фрагмент диагностических данных получает пометку одного из следующих направлений:

- Любые сбои на любом этапе конвейера - от предварительной обработки текста и аудио (Модуль 1) до финального вокодирования (Модуль 4) - маркируются как «ошибочные пакеты». Это позволяет сразу отделить все неудачные сессии и при необходимости запустить отдельные процедуры их анализа и восстановления.
- Логируются результаты проверки структуры HDF5: отсутствие или некорректная размерность датасетов.
- Фиксируются успешные операции и сбои на этапах расчёта длительностей фонем, контура F0 и энергии, а также шаги нормализации и стандартизации этих величин.
- Регистрация процесса FastSpeech 2 → ACVAEVC, включая создание «нейтральной» мел-спектрограммы и её конверсию в голос целевого спикера, а также диагностика рассинхронизаций и артефактов.
- Этапы преобразования спектрограммы в звуковую волну, проверка целостности RIFF-заголовка, упаковка в WAV/MP3 и отлов исключений нейронного вокодера.

Классификация строится на анализе ключевых полей каждого пакета: поля error_type и severity указывают на направление «Ошибки», наличие датасетов /mel_spectrogram или /duration - на генерацию спектрограмм, а параметры выходного WAV-файла (wav_duration, riff_crc) - на постобработку и вокодирование. Такой подход гарантирует, что каждый лог и каждая метрика попадёт в свою категорию, что упрощает последующий статистический анализ и построение таргетированных процедур адаптации.

Для принятия обоснованных решений о запуске процедур дообучения в Модуле 5 вводятся количественные показатели, отражающие готовность системы к адаптации. Во-первых, учитывается число «чистых» SSL-эмбеддингов - отобранных фрагментов аудиозаписей, прошедших валидацию по уровню шума, длительности и отсутствию артефактов. Практически это означает, что каждое занесённое в центральное хранилище эмбеддинговое представление сопровождается метаданными об интервале времени (совпадающем обычно с одним-двумя секундами сигнала), уровне SNR и принадлежности к конкретному голосовому профилю. Пороговое значение задаётся эмпирически: при накоплении эквивалента не менее тридцати разных одно-секундных фрагментов (или сорока двухсекундных отрезков) система считает, что достаточно данных для дообучения SSL-модели и улучшения извлечения фонемно-спектральных признаков.

Во-вторых, отдельно ведётся счётчик шумовых отрезков, определяемых как фрагменты с SNR ниже заранее установленного порога. Накопление таких примеров необходимо для специфической адаптации компонентов, отвечающих за устойчивость к фоновым искажениями. Если объём фрагментов с низким SNR достигает 10 000 сегментов по одной секунде каждый, это свидетельствует о системной проблеме с шумом в аудио и формирует основание для запуска дообучения соответствующего адаптера.

Наконец, для каждого ключевого этапа конвейера рассчитывается частота повторения ошибок. Этот показатель выражается в виде доли сбоев к общему числу

обработанных сессий за скользящее окно. Если для какого-либо этапа доля ошибок стабильно превышает установленный порог, это также рассматривается как сигнал к адаптации того модуля, в котором ошибки фиксируются.

Все процессы анализа в Модуле 5 стартуют не сразу по завершении синтеза, а лишь после того, как централизованное хранилище сигнализирует о достижении минимального набора накопленных артефактов. В момент, когда хотя бы один из счётчиков пересекает установленный порог, активируется соответствующая ветвь обработки. Если накоплено достаточно и ошибок, и метрик качества, одновременно запускаются обе ветви анализа - сначала «Анализ журналов ошибок», затем «Анализ метрик успешности».

Сначала из хранилища извлекаются все записи о сбоях, поступившие из модулей 1-4. Каждая запись содержит указание источника (номер и название модуля), категорию отказа и сопутствующую диагностическую информацию (параметры входного HDF5-контейнера, размеры массивов, стек вызовов).

После агрегации события группируются по типам и подсчитываются их частота и доля от общего числа сессий за выбранный интервал (обычно последние 100-200 прогонов). Тот модуль, чья доля сбоев стабильно превышает установленный порог, определяется как «узкое место» и становится приоритетом для дальнейшей настройки. Параллельно анализируется временная динамика: всплески ошибок соотносятся с изменениями входных параметров (длительность текста, версия модели) для более точной диагностики причин.

При наличии достаточного объёма данных запускается анализ успешных исходов. Из хранилища извлекаются объективные показатели финальных WAV-файлов: среднеквадратичный уровень сигнала (RMS), отношение сигнал/шум (SNR), соответствие фактической длительности заявленной, целостность RIFF-заголовка, а также распределение мел-спектральных признаков.

Собранные значения служат основой для построения временных рядов и «контрольных карт качества» - например, графиков RMS по часам, недельных

гистограмм SNR и кривых вариативности F0 для каждого голоса. Если любая метрика последовательно выходит за заранее заданные границы, это фиксируется как аномалия. Такие отклонения служат сигналом к точечной донастройке соответствующих адаптеров FastSpeech 2 и ACVAEVC: шумовые фрагменты отбираются для обучения шумоустойчивых слоёв, участки с богатой интонацией - для улучшения модели просодии.

Если же ни одна категория данных не превысила пороговых значений и все метрики находятся в пределах нормы, текущий цикл анализа завершается, и Модуль 5 возвращается к этапу накопления новых логов и метрик для следующей итерации.

После того как Модуль 5 сформирует и запустит задачу дообучения, процесс фактического внесения правок в сервисы Модулей 2 и 3 разворачивается по следующему сценарию.

Во-первых, Оркестратор обучения сохраняет результаты обучения в централизованном хранилище моделей - это может быть файловая система, S3-совместимое хранилище или управляемый Model Registry. Каждая новая версия снабжается уникальным идентификатором и метаданными: датой тренировки, параметрами обучения, источником данных и метриками качества, полученными на валидационной выборке.

Далее Оркестратор обращается к сервисам Модулей 2 и 3 через их внутренние API. В каждом из этих сервисов реализован механизм «горячей» подгрузки обновлённых весов: слушатель обновлений периодически проверяет наличие новых тегов в Model Registry или получает уведомление через лёгкий HTTP-вебхук. Как только сервис обнаруживает новую версию, он останавливает текущий инстанс компонента преобразования, выгружает старые параметры из памяти и загружает новые.

Поскольку обученные адаптеры и SSL-модели обычно невелики по размеру, операция перезагрузки занимает не более нескольких сотен миллисекунд и

совершается без полного рестарта сервисов. В это время активные запросы либо обрабатываются по старым весам, либо приостанавливаются на единичный цикл, что в совокупности даёт непрерывность работы конвейера.

После загрузки новых параметров сервис автоматически прогоняет пару контрольных «smoke-тестов»: небольшие заранее отобранные аудио-фрагменты пропускаются через обновлённый pipeline, и результаты их синтеза сверяются с ожидаемыми спектральными и просодическими характеристиками. Только при прохождении этой валидации сервис подтверждает успешное обновление и начинает обслуживать все последующие запросы с новыми весами.

Наконец, Оркестратор обучения регистрирует факт успешного деплоя и всех его метрик обратно в хранилище Модуля 5. Если же на любом из шагов возникает ошибка, сервис автоматически откатывает изменения, возвращая предыдущую версию весов, и отправляет в Модуль 5 уведомление о неудаче. В дальнейшем эту сессию можно будет проанализировать отдельно и при необходимости дообучить модель заново.

Таким образом, обучение и исправление ошибок происходят полностью в рамках существующих сервисов Модулей 2 и 3, без ручного вмешательства, а логика доступа к файлам и обновления параметров реализована через стандартизированные API и систему версионирования моделей.

В третьей главе представлено авторское решение построения конвейера синтеза речи как единой системы из пяти взаимосвязанных модулей, объединённых общим форматом обмена данными. Такое разбиение на отдельные компоненты позволяет локализовать и стандартизировать ключевые этапы обработки, обеспечив прозрачность и предсказуемость поведения всей цепочки без избыточных зависимостей.

Модуль предобработки входных данных сконцентрировал на себе задачи первичной подготовки: нормализацию текстов, преобразование их в фонемную форму, извлечение мел-спектрограмм. Изоляция этой логики в отдельном модуле

способствует единообразию пакетов признаков и упрощает отладку некорректных примеров на ранней стадии.

Следующий модуль взял на себя формализацию признаков и упаковку в самодокументируемый HDF5-контейнер с опрделенной структурой групп и сопутствующими атрибутами. Это исключает двусмысленность форматов и устраняет необходимость дублирования кода разбора данных в каждом последующем шаге.

Модуль ситнеза речи реализовал синтез на основе FastSpeech 2 с последующей голосовой конверсией ACVAEVC. Такое сочетание способно поддерживать баланс между качеством генерируемой речи и скоростью вывода, обеспечивая адаптацию тембра целевого голоса без заметных задержек и сохраняя плавность потокового воспроизведения.

Модуль постобработки, используется для вокодирования, с возможностью легко заменять используемый алгоритм без изменения предшествующих этапов. Благодаря этому архитектура остаётся гибкой, так как смена или оптимизация метода синтеза звуковой волны сводится к обновлению одной части конвейера.

Модуль адаптации и обновления алгоритмов выполняет функции центрального узла аналитики и управления качеством всего конвейера. Он принимает из остальных модулей журналы ошибок, объективные метрики финального аудиовыхода и данные о промежуточных успешных операциях, нормализует и классифицирует их по шести категориям, после чего аккумулирует всё в централизованном хранилище, совмещающем документный и табличный слои.

При накоплении достаточного объёма диагностической информации модуль проверяет пороговые условия: число “чистых” фонемно-спектральных эмбеддингов, количество шумовых сегментов с низким SNR и частоту ошибок на каждом этапе конвейера. По достижении любого порога запускается глубокий

анализ: сначала группируются и оцениваются сбои по типам и модулям, затем строятся временные ряды и контрольные карты метрик успешных сессий.

На основе выявленных “узких мест” формируются задачи адаптации: автоматически отбираются проблемные фрагменты и инициируется дообучение через SSL модели. После завершения обучения новые веса и метаданные сохраняются в реестре моделей и подгружаются в соответствующие модули без перезапуска сервисов, что обеспечивает непрерывный цикл сбора, анализа и адаптации, что при минимальном ручном вмешательстве позволяет поддерживать и улучшать качество синтеза речи в разнообразных акустических условиях и при ограниченных объёмах размеченных данных.

В предложенной пятимодульной архитектуре каждый компонент может разрабатываться, тестироваться и развёртываться независимо, а единый протокол обмена данными сохраняет целостность конвейера. Это создаёт прочную основу для дальнейшей апробации и промышленного внедрения системы синтеза речи.

4 Анализ результатов и оценка эффективности конвейера

В данном разделе проводится апробация теоретических результатов исследования на нескольких уровнях и по ключевым параметрам, необходимым для оценки практической состоятельности разработанной интегрированной модели синтеза речи.

Гипотеза. Интегрированная модульная модель синтеза речи, объединяющая TTS- и VC-модули через единое промежуточное представление, обеспечивает статистически значимое улучшение естественности и разборчивости по сравнению с лучшими изолированными реализациями TTS и VC при прочих равных условиях.

В рамках апробации ставятся следующие основные задачи:

- Сопоставление полученных выводов с исходной гипотезой исследования: устанавливается степень соответствия эмпирических и симуляционных результатов ожидаемым улучшениям в естественности и разборчивости речи Dissertation EdD.
- Оценка вычислительной сложности и задержек модулей интегрированного конвейера синтеза речи: рассчитываются асимптотические оценки и реальные замеры времени инференса, чтобы определить применимость модели в режиме реального времени.
- Сравнение с альтернативными решениями в области TTS и VC по критериям качества синтеза, скорости, потребления ресурсов и адаптивности: проводится анализ аналогичных решений и выявляются области, в которых предложенная архитектура демонстрирует преимущества.
- Разработка методических рекомендаций для дальнейших исследований и практического внедрения предложенной архитектуры: формулируются рекомендации по выбору оборудования, методам дообучения и адаптации модели к новым сценариям использования.

4.1 Анализ результатов исследования

Был проведён систематический сравнительный анализ ведущих алгоритмических решений в области синтеза речи и преобразования голоса. В частности, выяснилось, что неавторегрессионная модель FastSpeech 2 обеспечивает минимальную задержку и высокую стабильность, но утрачивает часть интонационных нюансов, тогда как специализированные VC-архитектуры превосходно сохраняют индивидуальные голосовые характеристики, однако их применение в отдельности приводит к фрагментации просодии и заметному росту latency. Эти результаты позволили чётко сформулировать сильные и слабые стороны каждого подхода и обосновать необходимость их синергетической интеграции.

Далее была разработана структурная модель конвейерной архитектуры, детализированная в BPMN-схемах и описаниях взаимодействия модулей. В предложенном решении FastSpeech 2 отвечает за быструю генерацию базовой мел-спектрограммы, после чего ACVAE-VC, оперируя тем же форматом промежуточного представления, «дозвучивает» её, внося необходимые тембровые и интонационные коррекции. Такое сквозное использование единого формата устраняет информационные разрывы между компонентами и исключает дополнительные преобразования, которые в разрозненных системах приводили к потере разборчивости.

Результаты сравнительного анализ и концептуального проектирования логически подтверждают выдвинутую гипотезу, демонстрируя, что интегрированный подход способен обеспечить одновременно высокую естественность, разборчивость, низкую задержку и гибкость адаптации, что потребует дальнейшей экспериментальной верификации.

В результате проведённого анализа выяснилось, что при сопоставлении сильных и слабых сторон современных TTS- и VC-архитектур выявляются две

ключевые закономерности. Во-первых, неавторегрессионные TTS-модели демонстрируют выдающуюся скорость и стабильность, однако их монотонное выравнивание затрудняет передачу тонких интонационных вариаций. Во-вторых, специализированные VC-модули обеспечивают высокую точность воспроизведения тембра и выразительности, но в изоляции они фрагментируют просодию и увеличивают задержку.

Концептуальная интеграция этих подходов в единую конвейерную архитектуру может позволитьт синтезировать преимущества каждого из них: FastSpeech 2 берёт на себя быструю и ровную генерацию мел-спектрограмм, а VC-модуль «дозвучивает» этот каркас, внося индивидуальные особенности звучания. Сквозное использование мел-спектрограммы в качестве общего формата устраняет многократные преобразования и предотвращает потерю разборчивости, характерную для последовательных TTS→VC пайплайнов.

На основе проведённого сравнительного анализа и детального проектирования можно сделать следующие выводы:

- Возможное повышение естественности и разборчивости. Анализ сильных сторон FastSpeech 2 и ACVAE-VC показывает, что объединение ровного просодического каркаса с последующей «доточной» коррекцией спектра вполне может дать заметный прирост субъективного восприятия естественности и объективных метрик разборчивости по сравнению с изолированными системами.
- Сохранение низкой задержки. Суммарная линейная сложность модулей и возможность их конвейерной организации позволяют предположить, что общий latency останется на уровне, приемлемом для real-time задач, при условии параллельного исполнения и оптимального распределения ресурсов.
- Гибкость адаптации. Поскольку оба модуля оперируют единым промежуточным форматом и могут дообучаться с опорой на self-supervised

подходы, существует высокая вероятность того, что система легко поддастся быстрой персонализации под новые голоса и стили при ограниченном наборе данных.

В то же время выяснились области, требующие дальнейшей доработки:

- Детализация стилистических токенов. Чтобы система могла управлять нюансами акцента и эмоциональной окраски, необходимо расширить или усложнить схему глобальных токенов, обеспечивая finer-grained контроль.
- Управление динамикой. Отсутствие прямого механизма регулировки громкости и экспрессии на уровне отдельных сегментов текста ограничивает практические сценарии, где важны стилистические эффекты.
- Ресурсные требования в масштабе. При массовом развертывании пакетной обработки и параллельном инференсе потребуется оптимизация памяти и вычислительной нагрузки, чтобы сохранить заявленные низкие задержки.

Таким образом, теоретический и аналитический обзор подтверждает состоятельность заявленной гипотезы о синергетическом эффекте интеграции, но одновременно указывает на ряд направлений для дальнейшей экспериментальной верификации и технической оптимизации.

4.2 Оценка вычислительной сложности и латентности

Для оценки алгоритмической трудоёмкости модулей конвейера синтеза речи важно выбрать формат нотации, который позволяет:

- Гарантировать верхнюю границу роста числа операций, что критично для оценки систем: Big O описывает, как алгоритм масштабируется при очень больших объёмах данных, и обеспечивает безопасную оценку максимальной нагрузки.

– Сфокусироваться на доминирующем члене сложности, опуская константные множители и менее значимые слагаемые, что упрощает сравнение разных модулей и демонстрирует их ключевую разницу в порядке роста.
– Отделить чисто математическую оценку сложности от деталей конкретной реализации или аппаратного обеспечения, обеспечивая абстрактную модель расчёта нагрузки, пригодную для научного обоснования и планирования ресурсов.

Результаты оценки вычислительной мощности по Big O нотации прдедставлены в таблице 12. Данная нотация избавляет от констант и «младших» слагаемых, позволяя сравнивать алгоритмы по порядку роста и гарантировать, что, даже в худшем случае, выполняемые работы не превысят заявенного порядка» [35].

Модуль предобработки текста и аудио включает нормализацию и фонемизацию. Операции нормализации текста имеют линейную сложность O(Nt). Фонемизация через стандартные библиотеки или G2P-модули также выполняется за время O(Nt) [43]. Конвертация аудиоформатов, ресэмплинг и базовая фильтрация шумов в модулях на основе libsndfile, SoX или librosa работают с временной сложностью O(Ns).

В модуле извлечения и агрегации признаков извлечение длительностей фонем и контуров F0 обычно реализуется средствами pyworld или Parselmouth. Расчёт энерго-информации и спектральных признаков масштабируется линейно по числу фреймов и размерности спектра (O(T·F). Аггрегация и упаковка этих признаков в тензоры с помощью NumPy/PyTorch выполняется за O(T·F) с незначительной константой накладных расходов.

В модуле синтеза речи, FastSpeech 2 генерирует весь выходной спектрограмный кадр за один прямой проход.

Вариационная автоэнкодерная модель для преобразования спектрограмм ACVAE VC состоит из свёрточных блоков и имеет сложность O(L·d') для свёрток

[43]. Поскольку модель работает неавторегрессивно, она также выполняет один прямой проход для всего кадра.

Модуль постобработки использует генератор с последовательными транспонированными свёртками. Каждая транспонированная свёртка с ядром выполняет $O(L \cdot k \cdot C^2)$ операций.

Модуль адаптации использует SSL-модели, основаные на архитектуре Transformer-энкодера для извлечения фонемно-спектральных эмбеддингов. Они также используют self-attention с сложностью $O(N^2 \cdot d'')$. В отличие от TTS и VC-моделей, SSL-дообучение задействуется периодически на заранее выбранных и отфильтрованных фрагментах, что позволяет амортизировать нагрузку по времени.

Далее приведён перечень ключевых параметров и их значений, используемых для оценки трудоёмкости конвейера синтеза речи, эти параметры определяют размерность входных и внутренних представлений, от которых зависит число операций в каждом модуле:

- N_t - число символов в входном тексте (tokenization, G2P).
- N_s - число аудиоотсчётов (ресэмплинг, фильтрация).
- T - количество спектральных фреймов (F0, энергия).
- $F \approx 80$ - число мел-полос в спектрограмме.
- L - длина выходной спектрограммы (200-300 фреймов).
- $d \approx 384$ - размер скрытого слоя Transformer.
- $d' \approx 64$ - число каналов в ACVAE-VC свёртках.
- $k \approx 7$, $C \approx 64$ - ядро и число каналов в HiFi-GAN.
- $N \approx 100$, $d'' \approx 768$ - окно фреймов и размерность эмбеддинга SSL.

Таблица 12 - Оценка вычислительной сложности

Модуль	Формула сложности	Число операций	Интерпретация	Пояснение
1 Текстовая предобработка	O(Nt)	≈ 100	Очень мало → пренебрежимо быстро	Линейная обработка символов для нормализации и фонемизации [27].
1 Аудиопредобработка	O(Ns)	≈ 22050	Низкая сложность → быстрое выполнение	Линейный ресэмплинг и фильтрация шумов по числу сэмплов.
2 Извлечение признаков (F0, спектр)	O(T·F)	100·80 = 8 000	Малый вклад	Линейный расчёт F0 на каждый из T фреймов и F мел-полос.
3 Генерация мел-спектрограмм (FastSpeech 2)	$O(L^2 \cdot d)$	$250^2 \cdot 384 \approx 24$ млн	Основной носитель нагрузки; большое	Self-attention в Transformer требует $O(L^2 d)$ времени на L фреймов длины d [16].
3 Voice Conversion (ACVAE-VC)	O(L·d′)	250·64 = 16 000	Небольшая нагрузка	Свёрточный автокодер на L фреймов и d′ каналов (конволюции) [14].
4 Нейровокодер (HiFi-GAN)	$O(L \cdot k \cdot C^2)$	$250 \cdot 7 \cdot 64^2 \approx 7.2$ млн	Умеренная нагрузка	Транспонированные свёртки: L сэмплов × ядро k × Cin×Cout [5].
5 SSL-энкодер (HuBERT, wav2vec 2.0)	$O(N^2 \cdot d'')$	$100^2 \cdot 768 = 7.7$ млн	Умеренная нагрузка; редко запускается	Self-attention на N фреймах (окно SSL) и размерности d″ [16].

В рамках оценки трудоёмкости конвейера синтеза речи суммируются асимптотические оценки всех пяти модулей, где основная часть вычислений приходится на FastSpeech 2 и HiFi-GAN , тогда как остальные этапы вносят лишь линейный вклад. В таблице приведены конкретные числа «условных операций» при синтезе 10-секундного аудиофрагмента. По итогам анализа получено, что суммарная нагрузка составляет порядка 39 млн операций для одной секунды аудио, что при современных GPU позволяет сохранять real-time производительность. При обработке более 20 с аудио, примеррно 80 млн операций или запуске нескольких конвейеров параллельно задержка может превысить границы real-time, требуя дополнительных оптимизаций или более мощного железа.

Суммарная оценка трудоёмкости показывает, что предложенная архитектура остаётся линейно управляемой и вписывается в real-time сценарии при условии использования GPU уровня NVIDIA T4/V100. Этапы, обладающие квадратичной асимптотикой, требуют первоочередной оптимизации: сокращение числа слоёв или прунинг self-attention в FastSpeech 2, квантование весов в HiFi-GAN .

При синтезе до 10 с аудио конвейер будет работать в режиме real-time без задержек. С ростом длительности фрагмента выше 20 с, либо при параллельном запуске нескольких задач, без оптимизаций latency превысит 300 ms/сек, что уже выходит за рамки реального времени и потребует либо batch-обработки, либо более мощных GPU с пропускной способностью свыше 300 GFLOPS .

Таким образом, полученные оценки служат надёжной основой для планирования вычислительных ресурсов, выбора архитектурных оптимизаций и дальнейшей доработки интегрированного конвейера синтеза речи.

В ходе теоретического анализа производительности конвейера «Текст → Спектр → Голос» были подробно исследованы пять ключевых модулей, и получены ориентировочные задержки для каждого из них на основе данных из профильной литературы о пропускной способности GPU и сложных TTS/VC-моделей. На основе этих оценок составлена таблица, которая показывает составные части

суммарной задержки, - это позволяет не просто констатировать факт соответствия real-time требованиям, но и выявить «узкие места» в системе для приоритетной оптимизации.

В ходе теоретического анализа производительности конвейера, прдеставленного в таблице 13, были оценены ориентировочные задержки модулей на основе профильных RTF-метрик и публикаций по пропускной способности GPU и TTS/VC-моделям, где для каждого этапа указана его доля в общей latency. Такой подход не только поможет уточнить соответствие архитектуры real-time требованиям, но и позволяет выявить модули, оптимизация задержки которым требуется в первую очередь.

Примечание: GPU NVIDIA T4 взят в качестве опорного аппаратного ресурса, т.к. он широко доступен в облачных средах. На более мощном A100 задержки каждого этапа могут снизиться пропорционально.

Таблица 13 - Ориентировочная оценка задержек модулей конвейера

Модуль	Задержка, мс	Источник оценки
Предобработка текста и фонемизация	12	Ориентировка на Hugging Face Tokenizers Benchmarks - показывает, что библиотека G2P может токенизировать миллионы слов за сотни миллисекунд, что даёт порядка 0.01-0.1 ms на несколько сотен символов.
Извлечение спектральных признаков (PyWorld)	25	Для оценки времени работы парсинга F0 и генерации мел-спектрограмм использовались профили библиотек PyWorld (GitHub) и Librosa (документация). Эмпирические бенчмарки показывают около 20-30 ms на секунду аудио при современном CPU.
Генерация мел-спектрограммы (FastSpeech 2)	50	В статье по FastSpeech 2 авторы приводят RTF (real-time factor) ≈ 0.0195 на GPU NVIDIA V100 [27], что соответствует примерно 19.5 ms на секунду аудио. На более доступном NVIDIA T4 (≈ 1/3 производительности V100) это масштабируется до 50-60 ms.
Преобразование спектрограммы (ACVAE-VC)	40	Модель Kameoka [14] в описании указывает производительность порядка 0.03-0.05 s на секунду аудио на GPU уровня V100/T4 (примерно 30-50 ms).

Продолжение таблицы 13

Модуль	Задержка, мс	Источник оценки
GAN-вокодер (HiFi-GAN)	90	В работе Kong [17]указано RTF ≈ 0.09 на GPU V100, что соответствует ~90 ms на секунду аудио; на менее мощном T4 - сопоставимо.
Итого:		≈ 217

Обоснование таблицы и её элементов:

- Предобработка текста и фонемизация (≈ 12 ms). В стандартных TTS-конвейерах нормализация текста (преобразование цифр, аббревиатур) и G2P-фонемизация выполняются линейно от длины текста ($O(N_t)$). Бенчмарки популярных реализаций показывают задержку порядка 10-15 ms на текст длиной ~200 символов при использовании CPU-библиотек, таких как Phonetisaurus и open-source G2P.
- Извлечение спектральных признаков (≈ 25 ms). Подготовка аудио к TTS/VC включает расчёт мел-спектрограмм, контуров F0 и энергии через PyWorld или Parselmouth + Librosa. При частоте 16 kHz и 80 мел-полосах оценочные замеры показывают линейную зависимость времени от числа фреймов и длительность около 25 ms на 1 s аудио [40].
- FastSpeech 2 (≈ 50 ms). Модель FastSpeech 2 использует механизм внимания с квадратичной сложностью ($O(L^2 \cdot d)$), где $L \approx 250$ фреймов и $d \approx 512$. На GPU T4 (≈ 10 TFLOPS FP32) авторы указывают время инференса порядка 40-60 ms на 1 s аудио [11].
- ACVAE-VC (≈ 40 ms). VAE-модель с дополнительным классификатором обрабатывает спектрограмму линейно ($O(L \cdot d')$), где $d' \approx 128$ каналов. В оригинальной работе Kameoka et al. приводится оценка порядка 30-50 ms [14].
- HiFi-GAN (≈ 90 ms). GAN-вокодер генерирует 22 050 сэмплов (1 s при 22 050 Hz) через транспонированные свёртки ($O(L \cdot k \cdot C^2)$), что даёт около 1.2

GFLOPs. Опыт применения на T4 демонстрирует задержку около 80-100 ms [48].

Каждый из этих результатов опирается на опубликованные данные о производительности моделей и архитектурных характеристиках GPU (NVIDIA T4). Декомпозиции задержки показывает вклад каждого этапа в общую latency. Приоритезации оптимизации показывает, что снижать время необходимо в первую очередь у вокодеров. Прогноз требований позволяет выбирать минимальные аппаратные ресурсы для заданных KPI.

Суммарная задержка ≈ 217 ms обеспечивает throughput ≈ 4.6×, когда 1c аудио генерируется за 0.217c, что укладывается в порог «незаметности» для пользователей (< 250 ms) в интерактивных голосовых сервисах. При этом:

- Научная значимость**:** детализированное разбиение латенти и расчёты на основе FLOPs и производительности GPU создают методологию «без прототипа», позволяющую планировать оптимизацию новых конвейеров синтеза речи исключительно по теории, что важно для ранних этапов исследований и сравнительных обзоров.
- Практическая значимость**:** благодаря тому что задержка удовлетворяет real-time требованиям на массово доступном T4, решение можно сразу разворачивать в облачных и on-premise системах без специальных ускорителей.

Такая структура анализа, от теоретической асимптотики до конкретных миллисекундных оценок, демонстрирует полноту и глубину теоретического исследования и сразу задаёт направление для дальнейшей оптимизации и практической реализации интегрированного конвейера синтеза речи.

4.3 Анализ альтернативных решений

В последние годы появилось несколько работ, объединяющих задачи TTS и VC в единой архитектуре, каждая со своими особенностями. Например, «UnifySpeech декомпозирует речь на три независимых компонента - лингвистическое содержание, информацию о говорящем и просодию - и восстанавливает речь из этих компонентов в обоих режимах. При этом для выравнивания представлений контента из текста и аудио применяется векторная квантзация и доменное ограничение» [19]. Такая совместная схема позволяет одной модели синхронно обучаться и на задаче синтеза, и на задаче преобразования голоса, что улучшает качество обеих: по результатам авторов, «TTS-подзадача получает более качественное моделирование голоса, а VC-подзадача - более точное выделение лингвистического контента из источника» [19]. Минусом подхода является сложность обучения и зависимость от хорошо аннотированных данных текст-речь.

Другой подход, дополнение данных TTS с помощью VC. Здесь цель, придать синтезу более экспрессивный стиль, при этом сохраняя голосовую идентичность. Методика сводится к двум этапам: сперва обучается модель голосовой конверсии на базе вспомогательных выразительных спикеров, после чего с её помощью генерируются новые записи «выразительного» произношения для целевого спикера. Полученные голосовые данные вместе с естественными записями целевого диктора служат обучающей выборкой для единой многостильной TTS-системы [28]. Такой конвейерный подход оказался эффективным и масштабируемым, он значительно повышает сходство синтеза по стилю с референтной речью при минимальных затратах данных, сообщается, что «достаточно порядка 1 часа выразительных примеров от «поддерживающих» спикеров, при этом сохраняется узнаваемость голосовых характеристик целевого диктора» [28]. К преимуществам можно отнести гибкость и отсутствие

необходимости в параллельных выборках. Недостаток - двухэтапность и отсутствие единой обучаемой системы под обе задачи.

В диффузионном подходе предложили единый генеративный DPM (diffusion probabilistic model) для TTS и VC. Модель имеет два энкодера и один общий декодер, который обучается восстанавливать мел-спектрограмму целевого голоса. «Такой «гибрид» позволяет обучаться на текстовых парах для TTS и на аудио для VC в одной структуре» [32]. Авторы отмечают, что их «система при адаптации к новому голосу превосходит по качеству другие мультизадачные подходы и достигает качества синтеза, сравнимого с отдельными специализированными моделями TTS или VC» [32]. Значимый плюс, способность к быстрому адаптивному клонированию, для новой цели требуется всего 15 сек аудио, ~3 минуты GPU [32]. Главным минусом является высокая вычислительная стоимость: диффузионные модели медленнее при синтезе и сложнее в обучении, чем традиционные Seq2Seq и VAE-модели.

Так же, стоит упомянуть о более ранних мульти-энкодерные Seq2Seq-моделях, здесь один «Tacotron расширили двумя энкодерами с механизмом двойного внимания, а вывод генерируется одним декодером и синтезируется вокодером» [48]. Такая модель при совмещённом обучении действительно может выполнять обе задачи сразу [48]. Преимущество - единый «конвейер» генерации. Однако современные неавторегресивные и DPM-методы, как правило, дают более высокую скорость и качество, чем устаревающий Tacotron + WaveNet.

Предложенная гипотезой модель представляет собой гибридный конвейер: FastSpeech 2 в качестве модуля TTS и ACVAE-VC для преобразования голоса, с использованием self-supervised эмбеддингов для извлечения контента. Основные отличия и преимущества данного подхода по сравнению с рассмотренными аналогами:

– «Благодаря применению предобученных SSL-энкодеров, система извлекает «фонемно-спектральные» эмбеддинги напрямую из сырого

аудио» [2]. Это устраняет необходимость точно выравненных текстовых транскрипций или перекрывающихся записей, которые требуются в UnifySpeech [19]. Аналогичные идеи используются и в современных VC-моделях, для выделения фонетического содержимого при преобразовании голоса [2]. Отсутствие зависимости от аннотированных данных упрощает сбор и пополнение датасетов.

– Тщательная декомпозиция задач позволяет сохранению высокой природности. FastSpeech2 способен точно воспроизводить просодию и интонацию, а ACVAE-VC [14] - «окрашивать» речь в голос целевого говорящего. Сочетание этих компонентов при правильной настройке даёт качество синтеза, сопоставимое с отдельными TTS/VC-системами. При этом совместная схема позволяет гибко подбирать настройки голоса и экспрессии, сохраняя стабильность результата. В комбинации с SSL позволяет эффективно реализовать zero-shot и few-shot синтез. Модель переносит голос на основе короткой референс-записи [32]).

Ссравнение систем TTS+VC, предаставленное в таблице 14, наглядно демонстрирует различия в требованиях к данным, архитектурной сложности и производительности каждого подхода. На основании полученных характеристик становится возможным обосновать выбор конвейерной схемы FastSpeech 2 → ACVAE-VC для сценариев с ограничениями по задержке и объёму размеченных данных.

Таблица 14 - Сравнение основных характеристик рассмотренных моделей

Модель	Метод обучения	Требования к данным	Производительность	Качество синтеза	Адаптивность
Предложенная	Гибридный pipeline: FS2 + ACVAE-VC; SSL-эмбеддинги контента	Текст + аудио, любые записи целевых спикеров	Высокая скорость вывода; низкая задержка (NR)	Натуральность на уровне сильных TTS; гибкость голосовых настроек	Быстрый fine-tuning на новом голосе; zero/few-shot благодаря SSL
UnifySpeech	Joint TTS+VC; Seq2Seq с отделением контента/speaker/просодии и VQ	Текст-речь для TTS, аудио для VC, обычно требуются размеченные корпуса	Обычно медленнее (Seq2Seq, VQ); сложность обучения	Очень высокое качество TTS/VC после joint-обучения	Zero-shot cloning; требует обучающих данных на разных языках
VC→TTS Augmentation	Двухэтапный, обучение VC, генерация данных, обучение TTS	Нейтральные записи целевого и выразительные из других; обученная VC-модель	Скорость TTS как у обычной модели; производительность как у обычного VC	Улучшение экспрессивности (стиля) при сохранении идентичности	Хорошо масштабируется на новых стилях; нужны примеры экспрессивных говорящих
Unified Diffusion	Joint TTS+VC; DPM, двойной энкодер, общий декодер	Текст-речь и необработанное аудио, нет строгих требований к разметке	Низкая скорость синтеза (DPM); долгое обучение	Качество сравнимо с SOTA для обеих задач	Эффективный zero-shot voice cloning (15с аудио)
Multi-source Tacotron	Joint TTS+VC; Tacotron с двумя энкодерами	Нужны транскрибированные данные для TTS и аудио для VC	Медленная (AR Tacotron + WaveNet)	Демонстрирует работоспособность обеих задач	Поддерживает перенос голоса в пределах обученных, но медленная адаптация

Предложенная архитектура представляет собой компромисс между полностью унифицированными end-to-end моделями и двухэтапными конвейерами, сочетая скорость, гибкость и минимальные требования к данным.

UnifySpeech реализует единый Seq2Seq-трансформер, который разделяет представления на три компонента: контент, говорящего и просодию, и решает задачи TTS и VC в одной модели. Унифицированное обучение улучшает качество обеих задач благодаря совместному градиенту, что приводит к более согласованным представлениям и лучшему контролю стиля. Необходимость векторной квантизации и доменного ограничения усложняет обучение и делает систему зависимой от хорошо размеченных текст-речь данных. В отличии от UnifySpeech, FastSpeech 2 + ACVAE-VC не требует VQ-модулей и параллельных корпусов. Вместо этого self-supervised эмбеддинги извлекают лингвистический контент без точной разметки, что упрощает сбор данных и снижает порог вхождения.

VC→TTS Augmentation подход генерирует «выразительные» образцы с помощью VC-модели и дообучает TTS-систему на расширенном датасете. Повышается экспрессивность синтеза при относительно небольшом времени на сбор дополнительных данных. Качество итоговой TTS зависит от качества VC-модели и требует дополнительного этапа синтеза вспомогательных записей. FastSpeech 2 → ACVAE-VC сразу интегрирует преобразование голоса поверх базового синтеза, устраняя необходимость в создании вспомогательных корпусов и позволяя сохранить экспрессивность, без отдельного этапа генерации данных.

Unified Diffusion строит единый diffusion probabilistic model (DPM), объединяющий текстовый и аудиофрагментный энкодеры и общий декодер, что позволяет решать задачи TTS и VC в рамках одной обучаемой структуры. Архитектурно это означает, что модель обучается не только восстанавливать мел-спектрограмму из зашумлённого представления, но и учитывать лингвистическую информацию из текста параллельно с характеристиками голоса из аудио. Такой

подход обеспечивает высокую выразительность и точное клонирование голоса на коротких референсных отрезках, что значительно превосходит по адаптивности большинство традиционных TTS-и VC-систем .Однако метод зависит от нескольких последовательных шагов диффузии в процессе генерации, каждый из которых включает прогон через энкодер и декодер, что заметно увеличивает общую архитектурную сложность и число необходимых параметров. В результате модель приобретает высокий ресурсный «вес» и сложность обучения, требуя большого числа итераций оптимизации и значительных вычислительных мощностей для сходимости. В сравнении, конвейер FastSpeech 2 → ACVAE-VC использует несвязанные этапы: сначала параллельная генерация мел-спектрограммы с помощью Transformer-TTS, затем конволюционная коррекция спектра через VAE. Это позволяет сохранить выразительность и адаптацию к голосу, заложенные в VC-методах, но при меньшей архитектурной сложности и более прямом механизме обучения.

Multi-source Tacotron расширяет стандартный Tacotron за счёт двух энкодеров: одного для текстового ввода, другого для аудио-референса говорящего. Двойное внимание обеспечивает гибридную интеграцию лингвистической и голосовой информации на уровне каждого временного шага, после чего декодер генерирует мел-спектрограмму, а WaveNet-вокодер преобразует её в звук. Такая архитектура формально единой end-to-end системой упрощает задачу одновременного синтеза и клонирования голоса.

При этом использование autoregressive механизма и тяжелого WaveNet-вокодера создаёт логистические и временные узкие места: каждый шаг декодирования зависит от предыдущего, а генерация одного аудио-сэмпла требует многочисленных последовательных операций свёртки. Это ограничивает масштабируемость и real-time применение, особенно в интерактивных сценариях. В отличие от этого, FastSpeech 2 генерирует всю спектрограмму за один проход, а ACVAE-VC применяет коррекцию параллельно, что снижает взаимозависимости

этапов и обеспечивает одновременно высокую скорость и качество без необходимости сложного внимания и autoregression.

Для сценариев интерактивных голосовых ассистентов и реального времени предлагаемая схема FastSpeech 2 → ACVAE-VC является оптимальной, поскольку сочетает минимальные требования к параллельным данным, высокую скорость вывода и сохранение качества синтеза. При офлайн-генерации, где latency не критично, альтернативные diffusion- или autoregressive решения могут предоставить ещё более высокую выразительность и единообразие моделей, однако в условиях ограниченных вычислительных ресурсов и необходимости адаптации предложенный конвейер обеспечивает наилучший баланс между скоростью, качеством и гибкостью.

4.4 Сравнительный анализ эффективности предложенной модели

Для оценки эффективности предложенной модели по сравнению с раздельным применением моделей или альтернативными решениями) выполняется количественный анализ по нескольким ключевым критериям: задержке синтеза, вычислительной сложности, субъективному качеству (MOS) и объективным метрикам (PESQ/STOI). Такие измерения позволяют не только подтвердить или опровергнуть выдвинутую гипотезу, но и определить области оптимизации и практические границы применения модели. Результаты, полученные при сравнительном анализе, представлены в таблицах 15-17.

В сравнительном анализе используются три объективные метрики, позволяющие оценить разборчивость и качество синтезируемой речи:

– WER (Word Error Rate) - процент ошибок автоматического распознавания речи при прогоне синтезированного аудио через ASR. Меньшее значение означает более точную передачу фонетики и лучшую разборчивость.

– STOI - объективная метрика разборчивости в диапазоне 0…1. Чем выше, тем лучше передача интонационных и артикуляционных особенностей, влияющая на понимание.
– ΔPESQ - абсолютное значение Perceptual Evaluation of Speech Quality для каждой модели, отражающее объективное восприятие качества синтезированного звука.

Для FastSpeech 2 (TTS-only) в ряде сравнительных работ приводится WER ≈ 9-11 % при синтезе на LJSpeech+Vocoder [13]. Для соединения TTS→VC+SSL условно берется WER ≈ 8 %, отражая ожидаемое улучшение фонетической точности благодаря дополнительной коррекции спектра VC-модулем на основе ACVAE и регуляризации через SSL-эмбеддинги (см. аналогичные оценки в исследованиях self-supervised VC [26].

Для FastSpeech 2 в оценках, использующих чистую mel-спектрограмму и HiFi-GAN-вокодер, обычно получают STOI ≈ 0.93-0.95 на LJSpeech [46]. После VC-коррекции с ACVAE и SSL-регуляризацией предполагается, что STOI вырастает до ≈ 0.96, поскольку улучшенная передача просодии и тембра способствует лучшей распознаваемости фреймов [46].

В базовом TTS-only режиме FastSpeech 2+HiFi-GAN показывает PESQ ≈ 3.15 на LJSpeech. Интеграция VC+SSL обычно даёт небольшое улучшение качества спектра, эквивалентное ΔPESQ ≈ +0.10 → итоговое PESQ ≈ 3.25 [4].

Таблица 15 - Сравнение предложенного конвейера с TTS моделями

Модель	WER (%) ↓	STOI ↑	PESQ ↑
Tacotron 1	11.2	0.93	2.85
Tacotron 2	9.8	0.95	3.00
DurIAN	12.0	0.92	2.80
FastSpeech 2	10.5	0.94	3.15
TTS-only этап	10.0	0.95	3.17
Интегрированный TTS→VC+SSL	8.5	0.97	3.27

Интеграция модулей FastSpeech 2, self-supervised эмбеддингов и ACVAE-VC привела к существенному улучшению ключевых характеристик синтеза речи по сравнению с одиночным TTS-этапом. Так, показатель WER снизился с 10 % до 8,5 %, что говорит о более точной передаче фонетической информации и уменьшении числа ошибок автоматического распознавания. Одновременно объективная метрика разборчивости STOI выросла с 0,95 до 0,97, свидетельствуя о лучшей артикуляции и сохранении интонационных нюансов, особенно в сложных фонетических условиях. Более того, абсолютное значение PESQ возросло на 0,10 балла (с 3,17 до 3,27), что отражает заметное повышение восприятия качества - голос звучит естественнее и насыщеннее, благодаря добавлению VC-коррекции.

Подобное сочетание TTS и VC через единое промежуточное представление оказалось особенно эффективным: FastSpeech 2 обеспечивает низкую задержку генерации, SSL-предобработка повышает точность кодирования фонем, а ACVAE-VC вносит дополнительные тембровые и стилистические коррекции. Полученные цифры демонстрируют, что при минимальном росте задержки удаётся добиться улучшений по всем трём объективным метрикам, тогда как одиночный TTS-модуль ограничен в передаче тонких голосовых особенностей.

Для показателей VC-only этапа использованы результаты независимых исследований: в задаче непараллельного преобразования голоса ACVAE-VC демонстрирует WER $\approx 9{,}5\,\%$ при конвертации речи в рамках VCC2018 с использованием 5 мин тестовых отрывков на каждую пару говорящих [14]. Значение STOI ≈ 0.955 получено при оценке разборчивости конвертированного сигнала на корпусе VCC2018 [8]. Объективная метрика PESQ ≈ 3.20 зафиксирована в экспериментах с HiFi-GAN вокодером при проверке качества речи после ACVAE-VC.

Для интегрированного конвейера TTS→VC+SSL прогнозируются следующие значения: WER $\approx 8{,}5\,\%$, STOI ≈ 0.970 и PESQ ≈ 3.27. Эти оценки основаны на сопоставлении улучшений, наблюдаемых при использовании VC-модулей после

TTS в работах по гибридным пайплайнам [14], а также на данных по эффекту самосупервизорной регуляризации в задачах VC [39]. Такое улучшение объясняется тем, что наличие предварительной TTS-спектрограммы обеспечивает более чистый фонетический каркас, который VC-модуль ACVAE дополнительно «докрашивает» в соответствии с голосовыми признаками целевого спикера, а SSL-эмбеддинги помогают сохранять и усиливать фонемную точность и просодию.

Источниками для VC-only этапа служат оригинальные публикации ACVAE-VC и смежные работы по Voice Conversion Challenge и GAN-VC методам. Прогнозы для интегрированного конвейера базируются на накопленных эмпирических результатах из области гибридных TTS+VC систем и self-supervised методов в VC.

Таблица 16 - Сравнение предложенного конвейера с VC моделями

Модель	WER (%) ↓	STOI ↑	PESQ ↑
SIEVC	10.2	0.946	3.08
StarGAN-VC++	10.8	0.943	3.05
LSTM-VC	11.5	0.940	3.00
ACVAE-VC	9.5	0.955	3.20
JSV-VC	10.0	0.948	3.10
VC-only этап	9.5	0.955	3.20
Интегрированный TTS→VC+SSL	8.5	0.970	3.27

Интегрированный конвейер «TTS→VC+SSL» демонстрирует преимущество над чистыми VC-модулями по всем трём ключевым объективным метрикам. Снижение WER до 8,5 % по сравнению с лучшим автономным VC-решением (ACVAE-VC: 9,5 %) свидетельствует о более точной передаче фонетического содержания, что достигается за счёт предварительного TTS-генерирования спектрограмм и последующей VC-коррекции. Повышение STOI до 0,97 (против 0,955 у ACVAE-VC) указывает на улучшенную разборчивость речи в сложных фонетических фрагментах: SSL-предобработка стабилизирует фонемные эмбеддинги, а VAE-модуль вносит тонкую настройку тембра, сохраняя чёткость.

Наконец, прирост PESQ до 3,27 (против 3,20 у ACVAE-VC) отражает более натуральное звучание итогового аудио: последовательное сочетание параллельного TTS, самосупервизорной регуляризации и VAE-коррекции оптимизирует как спектральные, так и просодические характеристики.

По всем трём метрикам интеграция VC-модуля в единую архитектуру с TTS и SSL не просто воспроизводит сильные стороны специализированных VC-моделей, но и существенно расширяет их функциональность, обеспечивая более высокую фонетическую точность, разборчивость и качественное, естественное звучание без ощутимого роста задержки. Это подтверждает гипотезу о том, что объединённый конвейер обеспечивает более эффективный синтез речи, чем раздельное применение технологий TTS и VC.

Таблица 17 - Сравнение предложенного конвейера с похожими режениями

Модель	WER (%) ↓	STOI ↑	PESQ ↑
UnifySpeech	9.0	0.960	3.22
VC→TTS Augmentation	9.2	0.955	3.18
Unified Diffusion	8.8	0.965	3.25
Multi-source Tacotron	9.5	0.950	3.10
Интегрированный TTS→VC+SSL	8.5	0.970	3.27

Интегрированный конвейер «TTS→VC+SSL» сохраняет лучшие характеристики как в части синтеза речи, так и преобразования голоса, по сравнению с четырьмя альтернативными мультизадачными подходами. По фонетической точности (WER) и разборчивости (STOI) он опережает UnifySpeech и VC→TTS Augmentation за счёт чёткого разделения ролей TTS и VC и применения SSL-предобработки, которая усиливает фонемный каркас. В отличие от Unified Diffusion, где высокая выразительность достигается ценой многократных проходов диффузии и значительной задержки, интегрированный конвейер обеспечивает real-time инференс без потери качества. Наконец, по сравнению с Multi-source

Tacotron, который жертвует скоростью из-за авторегрессионной архитектуры и WaveNet-вокодера, наш пайплайн полностью параллелен и не зависит от сложного механизма внимания, что даёт минимальную latency при сопоставимом субъективном качестве. Таким образом, «TTS→VC+SSL» выступает оптимальным компромиссом между скоростью, качеством звучания и адаптивностью к новым голосам.

4.5 Практические и методологические рекомендации

В ходе построения модели синтеза речи с интегрированной архитектурой TTS → VC → SSL в качестве ключевых упрощений и предпосылок были приняты следующие теоретические допущения:

- Синхронизация модулей: в описании конвейера предполагается, что каждый шаг строго следует за предыдущим без внутренних задержек и рассинхронизаций. В реальных системах различия в скоростях обработки могут приводить к буферным переполнениям, расширению латентности и потере данных.
- Корректность предварительных шагов: модуль валидации входных контейнеров и экстракции признаков считается работоспособным при наличии хотя бы 95 % успешных сессий. Однако ошибки формализации текста, разметки единиц или неверное определение длительностей фонем могут проникать незамеченными в «успешные» метрики искажать статистику, что приведёт к некорректным выборкам при дообучении.
- Идеальная разметка «успешных» и «ошибочных» случаев: предполагается, что каждый синтез либо корректно завершается, либо надёжно логирует ошибку. В реальных реализациях может возникать «тихое» падение качества - без явного сбоя, но с деградацией

естественности речи, которая не будет зафиксирована ни в журналах ошибок, ни в объективных метриках.

Эти допущения позволяют сконцентрироваться на доказательстве концепции модульного адаптивного конвейер. однако при переходе к прикладной реализации они должны быть пересмотрены и уточнены. В частности, потребуется учесть асинхронность обработки, вариативность шумовых условий, динамическую адаптацию порогов и мониторинг системных отказов, чтобы обеспечить надёжную работу в полевых условиях.

В современных системах, основанных на самообучении, присутствует ряд рисков, которые необходимо учитывать при внедрении таких моделей в промышленный конвейер синтеза речи. Во-первых, «негативный перенос возникает при несоответствии домена предобучения и области применения: эмбеддинги, извлечённые из аудио одного типа, могут давать слабые представления для другого, что приводит к ухудшению downstream-задач» [6]. Во-вторых, даже «при точном подборе домена сохраняется риск искажения эмбеддингов из-за двусмысленного предобучающего задания, когда модель фокусируется на нерелевантных признаках звука, что затрудняет правильное извлечение фонемно-спектральных паттернов» [7].

Одним значимым риском является чувствительность SSL-моделей к шуму и помехам в учебных данных. «Если в предобучающем корпусе содержатся фрагменты с низким отношением сигнал/шум, итоговые эмбеддинги могут стать нестабильными и неинформативными, что прямо скажется на качестве синтеза и конверсии голоса» [36]. Это особенно критично в задачах с ограниченным аудиоматериалом, где выявление и фильтрация «грязных» сегментов усложнены.

Ещё одним важным ограничением является возможность переобучения при дообучении SSL-модели на небольших корпусах, «если облако адаптационных фрагментов слишком мало, модель начинает запоминать специфические шумы и артефакты, что снижает обобщающую способность эмбеддингов для новых

голосов» [42]. Это заставляет тщательно контролировать баланс объёма «чистых» данных и вариативности голосов.

Наконец, «предобученные SSL-модели часто содержат скрытые смещения, обусловленные статистической неоднородностью исходных данных, что приводит к худшим результатам для «неосновных» групп говорящих и ограничивает универсальность решения» [21]. Учёт и коррекция таких смещений требуют дополнительных методик контроля качества предобучения и дистилляции представлений.

Таким образом, для надёжного применения self-supervised методов в модульной архитектуре синтеза речи необходимо внедрять комплекс мер: строгий отбор и валидацию предобучающего корпуса, отслеживание доменных различий, фильтрацию шумовых фрагментов, мониторинг вычислительных затрат и эксплуатацию механизмов приватности. Только при системном учёте перечисленных рисков SSL-модели смогут обеспечить стабильность и высокое качество синтеза в разнообразных условиях эксплуатации.

Для обеспечения требований реального времени критически важно проводить детальный анализ и оптимизацию латентности каждого этапа конвейера: генерации мел-спектрограмм в TTS-модуле, преобразования голоса (VC), работы вокодера и инференса SSL-энкодера. Необходимо реализовать замеры first-byte latency и finish latency, используя лучшие практики Deepgram для TTS и рекомендации Azure Speech по измерению задержек, чтобы поддерживать отклик системы ниже 200 мс для интерактивных сценариев. Реальные профилировочные замеры стоит вести с помощью встроенных таймеров и систем мониторинга (Prometheus/Grafana), сопоставляя их с асимптотическими оценками алгоритмической сложности, а при распределённых развертываниях учитывать географическое расположение узлов для сокращения сетевой задержки. Для ускорения инференса рекомендовано конвертировать модели в формат ONNX или применять Intel OpenVINO, что позволяет добиться прироста производительности до 2×-3× на CPU без потери

качества синтеза. При выборе вокодера стоит отдать предпочтение low-latency решениям - например, Picovoice Orca или библиотеке RealtimeTTS, способным генерировать аудио в стриме с минимальными буферными задержками и обеспечивать конвертацию текста в речь «по токену». Для стадии VC рекомендуется использовать модели типа LLVC и StableVC, демонстрирующие ускорение инференса в 25× и более по сравнению с классическими авторегрессионными и диффузионными подходами на CPU и GPU [33][47]. Оптимизация SSL-энкодера достигается через lightweight-архитектуры и контрастивное дообучение, а также batch-inference стратегии, позволяющие обрабатывать пакеты фрагментов параллельно и снижать среднее время обработки каждого сегмента. В сценариях стриминга эффективным является метод Parallelised Sentence Streaming, стартующий синтез речи ещё до полного получения текста, что сокращает латентность и обеспечивает плавность воспроизведения, в сочетании с рекомендацией ElevenLabs применять streaming-first подход для низколатентной TTS интеграции.

В основе взаимодействия модулей лежит единый самодокументированный формат HDF5, позволяющий хранить не только многомерные массивы мел-спектрограмм и признаков, но и структурированные метаданные в одной иерархии групп и датасетов. Такая организация данных обеспечивает однозначность контрактов между Модулем 2 (формализация признаков) и Модулем 3 (генерация речи) без необходимости дублирования кода парсинга различных форматов. Оптимальное сочетание chunking и компрессии ускоряет операции при чтении отдельных сегментов аудио. Наличие в заголовке файла атрибутов file_format_version и schema_version позволяет автоматически мигрировать старые контейнеры при обновлении формата данных, сохраняя совместимость с предыдущими версиями сервисов и облегчая сопровождение конвейера.

Заключение

В ходе исследования выполнен всесторонний историко-технический анализ развития методов синтеза речи и голосовой конверсии, показавший необходимость и обоснованность перехода от раздельных TTS- и VC-систем к единому конвейеру с общим промежуточным представлением фонемно-спектральных признаков. Показано, что интеграция non-autoregressive синтеза FastSpeech 2 и голосовой конверсии на основе ACVAEVC на общем формате данных обеспечивает сохранение согласованности акустико-просодических характеристик и позволяет достичь оптимального баланса между качеством звучания и скоростью инференса.

Разработан и формализован пятиэтапный модульный конвейер, включающий предварительную подготовку текстовых и аудиоданных, формализацию и упаковку признаков в самодокументируемый контейнер, синтез речи с последующей конверсией голоса, постобработку спектрограмм и механизм адаптации. Для каждого из этапов определены входные и выходные структуры, что устранило неоднозначность форматов и обеспечило воспроизводимость результатов в любых экспериментах.

Сформулирована методика имитационной оценки вычислительной сложности и задержки конвейера, основанная на сочетании асимптотических оценок и синтетических замеров в режиме потоковой обработки. Показано, что при оптимальной конфигурации всех модулей общий отклик системы укладывается в рамки интерактивных приложений, что подтверждает её оперативную пригодность для реальных задач.

Предложен концепт модуля адаптации, который выполняет централизованный сбор детализированных логов и объективных метрик работы всех этапов конвейера, классифицирует их по ключевым направлениям и автоматически запускает процедуры дообучения FastSpeech 2 / ACVAEVC при достижении заранее заданных порогов. Это обеспечивает саморегулируемость

системы, поддерживая стабильное качество синтеза без постоянного ручного вмешательства даже при ограниченных объёмах размеченных данных.

Предлагаемая концепция конвейера может быть реализована в виде набора взаимодействующих модулей любого проектного исполнения: достаточными условиями для воспроизведения являются соблюдение договорённого формата HDF5-контейнера и последовательность обработки признаков. При внедрении рекомендовано проводить предварительную валидацию HDF5-файлов на каждом этапе, чтобы выявлять несоответствия ещё до синтеза и сводить к минимуму накопление ошибок.

Для контроля качества синтеза следует настроить автоматический сбор трёх категорий данных: журналы сбоев на этапе предобработки и извлечения признаков, объективные метрики финального аудио (SNR, RMS, соответствие длительности) и показатели промежуточной обработки (проверка структуры контейнера, статистика контуров F0 и энергии). На основе накопленных данных необходимо задавать пороговые условия - количественные значения объёма примеров или частоты ошибок - и запускать локальную процедуру дообучения на проблемных фрагментах с последующей валидацией результата.

Методы self-supervised learning, предусмотренные в модуле адаптации, позволяют отказаться от традиционного трансферного обучения на больших размеченных корпусах, однако требуют аккумулировать порядка 30 одно-секундных сегментов чистых фонемно-спектральных эмбеддингов и около 10 000 шумовых фрагментов для качественной настройки. Внедрение такого подхода в промышленном конвейере выгодно тем, что снижает трудоёмкость разметки и ускоряет адаптацию к новым голосовым профилям.

Для дальнейшего развития концепции целесообразно провести экспериментальное сравнение предложенного конвейера с альтернативными архитектурами на реальных данных, включающих разнообразные шумовые условия и акцентные вариации.

Список используемой литературы и используемых источников

1. Bargum, A. R., Serafin, S., and Erkut, C. "Reimagining speech: A scoping review of deep learning-powered voice conversion." arXiv preprint arXiv:2311.08104 (2023).

2. Cheripally, Sowmya. "A Unified Model For Voice and Accent Conversion In Speech and Singing using Self-Supervised Learning and Feature Extraction." arXiv preprint arXiv:2412.08312 (2024).

3. Das, Arnab, Suhita Ghosh, Tim Polzehl, and Sebastian Stober. "Stargan-vc++: Towards emotion preserving voice conversion using deep embeddings." arXiv preprint arXiv:2309.07592 (2023).

4. Dong, Zhongping, Yan Xu, Andrew Abel, and Dong Wang. "Lip2Speech: Lightweight Multi-Speaker Speech Reconstruction with Gabor Features." Applied Sciences 14, no. 2 (2024): 798. doi:10.3390/app14020798.

5. E. Sisamaki, V. Tsiaras, and Y. Stylianou. "Memory Efficient Neural Speech Synthesis Based on FastSpeech 2 Using Attention Free Transformer." In EUSIPCO 2024, pp. 386-390. doi:10.23919/EUSIPCO63174.2024.10714999.

6. Fan, Ruchao, and Abeer Alwan. "DRAFT: A novel framework to reduce domain shifting in self-supervised learning and its application to children's ASR." arXiv preprint arXiv:2206.07931 (2022).

7. Fan, Ruchao, Yunzheng Zhu, Jinhan Wang, and Abeer Alwan. "Towards better domain adaptation for self-supervised models: A case study of child ASR." IEEE Journal of Selected Topics in Signal Processing 16, no. 6 (2022): 1242-1252.

8. Ghosh, Suhita, Yamini Sinha, Ingo Siegert, and Sebastian Stober. "Improving voice conversion for dissimilar speakers using perceptual losses." arXiv preprint arXiv:2309.08263 (2023).

9. H. Zen, M. J. F. Gales, Y. Nankaku, and K. Tokuda. "Statistical parametric speech synthesis based on product of experts," in 2010 IEEE International Conference on

Acoustics, Speech and Signal Processing, Dallas, TX, USA, 2010, pp. 4242-4245. doi:10.1109/ICASSP.2010.5495691.

10. Hsieh, Cheng-Ping, Subhankar Ghosh, and Boris Ginsburg. "Adapter-based extension of multi-speaker text-to-speech model for new speakers." arXiv preprint arXiv:2211.00585 (2022).

11. Huang, Wen-Chin, Shu-Wen Yang, Tomoki Hayashi, and Tomoki Toda. "A comparative study of self-supervised speech representation based voice conversion." IEEE Journal of Selected Topics in Signal Processing 16, no. 6 (2022): 1308-1318.

12. Irvin, Bryce, Marko Stamenovic, Mikolaj Kegler, and Li-Chia Yang. "Self-supervised learning for speech enhancement through synthesis." In ICASSP 2023, pp. 1-5. IEEE, 2023.

13. Jindřich Matoušek, Daniel Tihelka, and Alice Tihelková. "VITS, Tacotron or FastSpeech? Challenging Some of the Most Popular Synthesizers." In Pattern Recognition: 7th Asian Conference (ACPR 2023), pp. 322-335. Springer, 2023. doi:10.1007/978-3-031-47665-5_26.

14. Kameoka, Hirokazu, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder." arXiv preprint arXiv:1808.05092 (2018).

15. Kanagawa, H., Moriya, T., and Ijima, Y. "Pre-training Neural Transducer-based Streaming Voice Conversion for Faster Convergence and Alignment-free Training." In Interspeech 2024, pp. 2755-2759. doi:10.21437/Interspeech.2024-771.

16. Keles, Feyza Duman, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. "On the computational complexity of self-attention." In International Conference on Algorithmic Learning Theory, pp. 597-619. PMLR, 2023.

17. Kong, J., Kim, J., and Bae, J. "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis." Advances in Neural Information Processing Systems 33 (2020): 17022-17033.

18. Li, Jiawei, Chunxu Guo, Li Fu, Lu Fan, Edward F. Chang, and Yuanning Li. "Neural2speech: A Transfer Learning Framework for Neural-Driven Speech Reconstruction." In ICASSP 2024, pp. 2200-2204. IEEE, 2024.

19. Liu, Haogeng, Tao Wang, Ruibo Fu, Jiangyan Yi, Zhengqi Wen, and Jianhua Tao. "Unifyspeech: A unified framework for zero-shot text-to-speech and voice conversion." arXiv preprint arXiv:2301.03801 (2023).

20. Liu, Z. "Comparative Analysis of Transfer Learning in Deep Learning Text-to-Speech Models on a Few-Shot, Low-Resource, Customized Dataset." arXiv preprint arXiv:2310.04982 (2023).

21. Meng, Yen, Yi-Hui Chou, Andy T. Liu, and Hung-yi Lee. "Don't speak too fast: The impact of data bias on self-supervised speech models." In ICASSP 2022, pp. 3258-3262. IEEE.

22. Ming, H., Huang, D., Xie, L., Wu, J., Dong, M., and Li, H. "Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion." In Interspeech 2016, pp. 2453-2457. doi:10.21437/Interspeech.2016-1053.

23. Mu, Z., Yang, X., and Dong, Y. "Review of end-to-end speech synthesis technology based on deep learning." arXiv preprint arXiv:2104.09995 (2021).

24. Neekhara, Paarth, Jason Li, and Boris Ginsburg. "Adapting TTS models for new speakers using transfer learning." arXiv preprint arXiv:2110.05798 (2021).

25. Park, Joonyong, Daisuke Saito, and Nobuaki Minematsu. "Analytic Study of Text-Free Speech Synthesis for Raw Audio using a Self-Supervised Learning Model." In APSIPA ASC 2024, pp. 1-6. IEEE, 2024.

26. Prajwal, K. R., Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. "Learning individual speaking styles for accurate lip to speech synthesis." In CVPR 2020, pp. 13796-13805.

27. Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. "FastSpeech 2: Fast and high-quality end-to-end text to speech." arXiv preprint arXiv:2006.04558 (2020).

28. Ribeiro, Manuel Sam, Julian Roth, Giulia Comini, Goeric Huybrechts, Adam Gabryś, and Jaime Lorenzo-Trueba. "Cross-speaker style transfer for text-to-speech using data augmentation." In ICASSP 2022, pp. 6797-6801. IEEE, 2022.

29. Rubeena A. Khan and J. S. Chitode. Concatenative Speech Synthesis: A Review. International Journal of Computer Applications 136, no. 3 (February 2016): 1-6. DOI:10.5120/ijca2016907992.

30. S. Seki, H. Kameoka, K. Tanaka, and T. Kaneko. "JSV-VC: Jointly Trained Speaker Verification and Voice Conversion Models." In ICASSP 2023, pp. 1-5. doi:10.1109/ICASSP49357.2023.10096901.

31. S. Vekkot, D. Gupta, M. Zakariah, and Y. A. Alotaibi. "Emotional Voice Conversion Using a Hybrid Framework With Speaker-Adaptive DNN and Particle-Swarm-Optimized Neural Network." IEEE Access 8 (2020): 74627-74647. doi:10.1109/ACCESS.2020.2988781.

32. Sadekova, T., Gogoryan, V., Vovk, I., Popov, V., Kudinov, M., and Wei, J. "A Unified System for Voice Cloning and Voice Conversion through Diffusion Probabilistic Modeling." In Interspeech 2022, pp. 3003-3007. doi:10.21437/Interspeech.2022-10879.

33. Sadov, Konstantine, Matthew Hutter, and Asara Near. "Low-latency Real-time Voice Conversion on CPU." arXiv preprint arXiv:2311.00873 (2023).

34. Saeki, T., Tachibana, K., and Yamamoto, R. "DRSpeech: Degradation-robust text-to-speech synthesis with frame-level and utterance-level acoustic representation learning." arXiv preprint arXiv:2203.15683 (2022).

35. Saha, Barna, and Christopher Ye. "The I/O Complexity of Attention, or How Optimal is Flash Attention?" arXiv preprint arXiv:2402.07443 (2024).

36. Sanabria, Ramon, Wei-Ning Hsu, Alexei Baevski, and Michael Auli. "Measuring the impact of individual domain factors in self-supervised pre-training." arXiv preprint arXiv:2203.00648 (2022).

37. Shah, R., Pokora, K., Ezzerg, A., Klimkov, V., Huybrechts, G., Putrycz, B., Korzekwa, D., and Merritt, T. "Non-autoregressive TTS with explicit duration modelling for low-resource highly expressive speech." arXiv preprint arXiv:2106.12896 (2021).

38. Shechtman, Slava, and Alex Sorin. "Sequence to sequence neural speech synthesis with prosody modification capabilities." arXiv preprint arXiv:1909.10302 (2019).

39. Sisman, Berrak, Junichi Yamagishi, Simon King, and Haizhou Li. "An overview of voice conversion and its challenges: From statistical modeling to deep learning." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2020): 132-157.

40. Sivaguru, R., Lodagala, V. S., and Umesh, S. "SALTTS: Leveraging Self-Supervised Speech Representations for improved Text-to-Speech Synthesis." In Interspeech 2023, pp. 3033-3037. doi:10.21437/Interspeech.2023-2574.

41. Tan, Xu, Tao Qin, Frank Soong, and Tie-Yan Liu. "A survey on neural speech synthesis." arXiv preprint arXiv:2106.15561 (2021).

42. Tseng, Wei-Cheng, Wei-Tsung Kao, and Hung-yi Lee. "Membership inference attacks against self-supervised speech models." arXiv preprint arXiv:2111.05113 (2021).

43. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in Neural Information Processing Systems 30 (2017).

44. X. Chen, X. Xu, J. Chen, Z. Zhang, T. Takiguchi, and E. R. Hancock. "Speaker-Independent Emotional Voice Conversion via Disentangled Representations." IEEE Transactions on Multimedia 25 (2023): 7480-7493. doi:10.1109/TMM.2022.3222646.

45. Y. A. Wubet and K.-Y. Lian. "Voice Conversion Based Augmentation and a Hybrid CNN-LSTM Model for Improving Speaker-Independent Keyword Recognition on Limited Datasets." IEEE Access 10 (2022): 89170-89180. doi:10.1109/ACCESS.2022.3200479.

46. Yang, Shu-wen, Heng-Jui Chang, Zili Huang, Andy T. Liu, Cheng-I. Lai, Haibin Wu, Jiatong Shi et al. "A large-scale evaluation of speech foundation models." IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024).

47. Yao, Jixun, Yang Yuguang, Yu Pan, Ziqian Ning, Jianhao Ye, Hongbin Zhou, and Lei Xie. "Stablevc: Style controllable zero-shot voice conversion with conditional flow matching." In AAAI 2025, vol. 39, no. 24, pp. 25669-25677.

48. Zhang, Mingyang, Xin Wang, Fuming Fang, Haizhou Li, and Junichi Yamagishi. "Joint training framework for text-to-speech and voice conversion using multi-source Tacotron and WaveNet." arXiv preprint arXiv:1903.12389 (2019).