

Автоматизированное формирование образовательных видеоматериалов на основе модульной системы цифрового двойника преподавателя

А.Г. Леонов^{1, 2, 3, 4✉}, К.А. Машенко^{1, 2, 3}, Н.С. Мартынов²,
К.К. Пчелин², А.В. Шляхов²

¹ МГУ имени М.В. Ломоносова, г. Москва, 119991, Россия

² НИЦ «Курчатовский институт» – НИИСИ, г. Москва, 117218, Россия

³ Московский педагогический государственный университет, г. Москва, 119991, Россия

⁴ Государственный университет управления, г. Москва, 109542, Россия

Ссылка для цитирования

Леонов А.Г., Машенко К.А., Мартынов Н.С., Пчелин К.К., Шляхов А.В. Автоматизированное формирование образовательных видеоматериалов на основе модульной системы цифрового двойника преподавателя // Программные продукты и системы. 2025. Т. 38. № 4. С. 694–703. doi: 10.15827/0236-235X.152.694-703

Информация о статье

Группа специальностей ВАК: 2.3.5, 2.3.8

Поступила в редакцию: 20.06.2025

После доработки: 15.07.2025

Принята к публикации: 24.07.2025

Аннотация. В статье представлена система персонального цифрового двойника преподавателя для автоматизированного формирования и актуализации обучающих видеоматериалов в цифровой образовательной платформе. Предложенная архитектура системы состоит из набора взаимосвязанных модулей, каждый из которых решает специализированную задачу и обеспечивает гибкость и масштабируемость решения. На первом этапе автоматическое распознавание речи реализовано с использованием открытой мультязычной модели Whisper. Это позволяет обрабатывать исходные аудиоданные без предварительного дообучения под конкретный домен данных. Далее извлеченный текст обрабатывается универсальным препроцессором, обеспечивающим нормализацию числительных, предсказание ударений в словах и транскрипцию в международный фонетический алфавит. Для синтеза речи применяется модифицированная версия архитектуры CoquiTTS, поддерживающая zero-shot клонирование голоса по короткому образцу речи преподавателя. Формирование визуальной составляющей обеспечивается модулем LivePortrait, генерирующим реалистичный видеопортрет преподавателя по предоставленной им фотографии. Синхронизация движений губ – синтезированным аудиомодулем LipSync, что повышает воспринимаемое качество видео. Завершающим этапом является автоматическая сборка презентации с интеграцией полученных системой видеофрагментов. В работе проведен сравнительный анализ альтернативных архитектур для каждого модуля, выделены критерии оценки качества синтезированной речи и видеогенерации, а также описаны особенности разработки русскоязычного текстового препроцессора. Экспериментальные результаты показывают сокращение трудозатрат преподавателя на создание и обновление обучающих материалов более чем на 70 % при сохранении их высокого уровня естественности. Обсуждены ограничения текстовой обработки формул и сокращений, а также рассмотрены перспективы оптимизации скорости работы модуля LipSync.

Ключевые слова: цифровой двойник, текстовый препроцессор, образовательные видеоматериалы, ASR, TTS, LipSync, zero-shot синтез голоса, русскоязычные TTS

Благодарности. Работа выполнена в рамках государственного задания НИЦ «Курчатовский институт» – НИИСИ по теме № FNEF-2024-0001 (1023032100070-3-1.2.1)

Введение. Разработка обучающих видеоматериалов преподавателями для цифровых образовательных платформ сопряжена со значительными трудозатратами. Одна часть проблем связана с технологическими и инфраструктурными требованиями, например, обеспечение профессионального качества аудиозаписи (отсутствие фоновых акустических шумов) подразумевает соответствующее оборудование и условия записи. Другая часть обусловлена сложностью фрагментарной модификации ранее созданного учебного контента. Необходимость постоянной актуализации и обновления материалов, вызванная изменением уровня подготовки целевой аудитории, непрерывным развитием педагогических методик и прогрессом

в соответствующей области знаний, как правило, требует полной перезаписи лекций, что является трудозатратной операцией. Таким образом, автоматизация процесса создания и (или) обновления учебных видеоматериалов является актуальной задачей.

Предложено использовать методику персонального цифрового двойника (аватара) преподавателя, способного воспроизводить текст образовательных материалов, произнесенный голосом педагога, и автоматически генерировать актуальные видеoverсии лекций. Важной задачей при этом становится поиск и анализ современных программных систем с открытым исходным кодом в этой области, поскольку только они обеспечивают гибкость разработки

и надежность использования без зависимости от внешних сервисов. Использование API коммерческих поставщиков (например, HeyGen) не гарантирует стабильности и вносит дополнительные риски при интеграции в цифровую образовательную платформу.

В данной работе предложена и реализована модульная архитектура системы персонального цифрового двойника преподавателя. Модульность подхода обеспечивает масштабируемость и поддержку актуальности системы за счет отсутствия зависимостей между различными компонентами при их доработке или замене. Выделено несколько ключевых компонентов: модуль автоматического распознавания речи (*Automatic Speech Recognition, ASR*), универсальный текстовый препроцессор, модуль синтеза речи (*Text to Speech, TTS*) с поддержкой zero-shot голосового клонирования, модуль генерации видеопортрета преподавателя из фотографии, LipSync-модуль для синхронизации мимики с синтезированным аудио, а также компонент автоматической сборки учебной презентации. Описаны архитектура, входные и выходные форматы данных, взаимодействие между модулями и логика конвейерной обработки информации.

Научная новизна исследования заключается в разработке модели интеграции открытых модулей автоматического распознавания речи, интеллектуальной текстовой предобработки, синтеза речи с возможностью zero-shot голосового клонирования, генерации видеопортрета и артикуляционной синхронизации в единый воспроизводимый конвейер, адаптированный под специфику русскоязычного образовательного контента. Предложены методы и алгоритмы, ключевым из которых является универсальный текстовый препроцессор, объединяющий нормализацию числительных, автоматическую расстановку ударений и транскрипцию в международный фонетический алфавит (*International Phonetic Alphabet, IPA*) в рамках единого модуля, что обеспечивает стабильность работы TTS-системы и высокую артикуляционную согласованность при LipSync-синхронизации. Обоснован выбор архитектур XTTS, LivePortrait и LatentSync на основе анализа качества генерации и затрат вычислительных ресурсов. Разработана программная архитектура системы с использованием модульной структуры, допускающей независимую замену и доработку компонентов без потери совместимости. Такой подход демонстрирует достижение уровня качества, сопоставимого с коммерческими плат-

формами, при сохранении контроля над архитектурой и минимизации зависимости от сторонних сервисов.

При реализации решения системы персонального цифрового двойника были использованы передовые методы. В качестве первых эффективных подходов в области ASR применялись статистические архитектуры на основе скрытых марковских моделей, часто в сочетании с гауссовыми смесевыми моделями. Они имели преимущество естественного временного моделирования структуры речи, статистической обоснованности и модульности (разделение акустических и языковых моделей). С появлением глубокого обучения область ASR кардинально преобразилась: глубокие нейронные сети показали превосходство над традиционными методами, особенно в акустическом моделировании [1]. Впоследствии были внедрены механизмы внимания и архитектуры трансформеров, что позволило повысить качество распознавания за счет учета долгосрочных зависимостей в последовательности аудио [2]. Однако трансформеры оказались чувствительны к разреженным входным данным (например, к отрезкам тишины) [3]. Для решения проблемы использовались разные способы, в частности, постобработка выходов трансформера с помощью BERT-подобных моделей [4] или регуляризация внимания с учетом информации о паузах [5].

Зарубежные исследователи представили мультязычную модель Whisper, обученную на большом размеченном аудиокорпусе объемом 680 тыс. часов (<https://github.com/openai/whisper/>). Она демонстрирует высокую производительность распознавания речи в реальном времени без необходимости дообучения под конкретный домен, а открытый исходный код и доступные веса модели позволяют выбрать ее в качестве базовой архитектуры ASR-модуля в данной работе.

В случае наличия только одного статического изображения преподавателя предусмотрен модуль генерации реалистичного видео с естественными движениями лица (видеопортрета). Ранее эта задача решалась с помощью диффузионных моделей генерации [6], однако такие подходы были вычислительно затратными и плохо управляемыми. Современные решения на основе выявления ключевых точек лица с последующей интерполяцией (фреймворки Keypoint) позволили перейти к более эффективному синтезу движений [7]. Одним из таких решений является модель LivePortrait

(<https://github.com/KwaiVGI/LivePortrait/>), обученная на большом объеме данных и способная генерировать естественно выглядящий видеопортрет за время, сопоставимое с работой ASR-модуля.

Задача преобразования текста в речь (TTS) также широко исследована. Первые системы синтеза речи использовали статистические парадигмы (например, НММ) [8]. Далее появились нейросетевые модели: Tacotron 2 с рекуррентными сетями и вниманием (<https://github.com/NVIDIA/tacotron2/>), полностью сверточная Deep Voice 3 [9], модели с трансформерами типа FastSpeech 2 (<https://speechresearch.github.io/fastspeech2/>). Среди современных многоязычных решений стоит отметить VALL-E, использующую автоэнкодер для акустических токенов [10]. Комбинированные подходы, сочетающие трансформеры и диффузии (Mars5 – система TTS, основанная на трансформерах и диффузионных моделях (<https://www.camb.ai/>), XTTS [11]), показали выдающиеся результаты. Для поддержки zero-shot голосового клонирования и мультилингвальности в данном проекте была выбрана открытая архитектура XTTS (реализация CoquiTTS) с IPA-предобработкой текста.

Все современные TTS-системы сталкиваются со сложной задачей обработки омографов, неявно указываемых ударений и специальных сокращений. Решение заключается в фокусе на предобработку текста (вплоть до фонетической транскрипции): многие системы для западных языков используют фонетические словари (например, CMU Dictionary) (<http://www.speech.cs.cmu.edu/cgi-bin/cmudict/>) и правила для обработки чисел и сокращений. Однако для русского языка готовые комплексные решения отсутствуют: открытые фонетические словари (RuStress – модуль для прогнозирования ударения в русских словах (<https://github.com/AIshutin/rustress>)) покрывают лишь базовые случаи, а обработка числительных и сокращений часто выполняется с помощью статичных скриптов. Отсутствие объемных размеченных данных и специализированных моделей влечет за собой синхронизацию нескольких узкоспециализированных инструментов, что снижает обобщаемость и качество системы.

Так, ключевое отличие решений для английского и русского языков заключается в уровне интеграции модулей: зарубежные разработки объединяют статические правила с гибкими ML-моделями в единый конвейер, тогда как на русском языке большинство систем представ-

ляет собой набор разрозненных инструментов. Это подчеркивает актуальность создания универсального препроцессора, способного стандартизировать обработку русскоязычных текстов без потери гибкости.

Синхронизация движений губ (LipSync) между сгенерированным TTS-системой аудио- и видеорядом является еще одним ключевым компонентом цифрового аватара, напрямую влияющим на качество визуальной интерпретации результата. Ранее наиболее распространенными решениями в этой области были модели на базе Wav2Lip [12], использующие пок кадровую генерацию изображения на основе краткосрочного контекста аудио. Повышение визуального качества в таких системах часто достигалось через дообучение и применение дополнительных реставрационных моделей, например, GFPGAN для коррекции артефактов лица [13]. Однако современные подходы стремятся улучшить как естественность мимики, так и стабильность кадров за счет более глубокого анализа аудиопотока и адаптации к движению губ. Среди таких архитектур можно выделить TalkLip (<https://github.com/Sxjdwang/TalkLip/>), DiffDub (<https://github.com/liutaocode/DiffDub/>) и KeySync (<https://antonibigata.github.io/KeySync/>), которые внедряют attention-механизмы или предобученные аудио-видео эмбединги для более точной синхронизации. Последние разработки, такие как LatentSync (<https://github.com/byte-dance/LatentSync/>), работают в скрытом пространстве признаков и демонстрируют наилучшие результаты по метрике LipScore [14] на открытых и экспериментальных датасетах, особенно в рамках решения задачи кросс-синхронизации. Тем не менее практически все актуальные модели LipSync продолжают использовать пок кадровый режим генерации, что исключает возможность параллельной обработки и делает скорость их работы критически зависимой от длины видео. Это требует компромиссного решения, основанного на балансе между вычислительной эффективностью и визуальной точностью.

Методология и стратегия авторского подхода

Предложенная система реализует конвейер обработки данных. На вход подаются фотографии преподавателя или короткие аудиозаписи его голоса, файл презентации лекции (например, в формате PowerPoint), который может содержать уже записанные ранее лектором видео-

фрагменты и текстовые слайды обучающего материала. При необходимости используются следующие этапы обработки.

Извлечение текста из лекционного материала производится с помощью модуля ASR и встроенных в презентацию заметок. Если в ней уже присутствует расшифровка слайдов, то она может использоваться напрямую. В противном случае ASR-модуль распознает и переводит речь из имеющихся аудио- или видеофрагментов в текстовую репрезентацию. На вход модуля поступает аудиопоток, на выходе формируется текст. В данной работе за ASR отвечает модель Whisper, выбор которой обусловлен наличием мультязычной поддержки, высокой точностью распознавания в условиях присутствия окружающего шума и возможностью работы в реальном времени в силу высокой вычислительной производительности.

Распознанный текст предоставляется преподавателю для интерактивного редактирования, после чего он становится основой для актуализации содержания и последующей генерации обновленной видеолекции.

Отредактированный текст подается на вход *текстового препроцессора*, который решает ключевые задачи подготовки текста для качественного синтеза речи. В задачи препроцессора входят преобразование числительных в полную текстовую форму, расшифровка аббревиатур и сокращений, корректная расстановка ударений в омографах, приведение математических формул к читабельному виду.

Основные проблемы, решаемые препроцессором, также включают некорректное произношение ударений в словах, неправильную интонацию синтезируемого голоса и ошибочное чтение числительных, аббревиатур и дат. Для их решения были рассмотрены два жизнеспособных подхода: адаптация компактной модели ruT5 [15] через дообучение на синтетических данных и использование большой языковой модели (*Large Language Model*, LLM) с настроенными промптами. Они пригодны для разных условий: при ограниченных вычислительных ресурсах предпочтительнее ruT5 (с учетом затрат на генерацию и обновление датасета), тогда как LLM обеспечивает гибкость и легкость адаптации к новым доменам. В рамках данного проекта выбран подход на основе модели Qwen-32B:q6 (<https://github.com/QwenLM/Qwen3/>) с оптимизированными промптами. Он позволил быстро настраивать правила преобразования (например, добавлять обработку математических формул или техниче-

ских аббревиатур) без переобучения. Для повышения робастности обработки реализована следующая гибридная схема: LLM осуществляет основную обработку текста, а узкоспециализированная дообученная модель (ruT5-base) корректирует особо сложные конструкции.

Для обучения ruT5-base был сформирован синтетический датасет размеченных пар «оригинальный текст – скорректированный текст» объемом более 720 тыс. примеров. Синтетические примеры генерировались с помощью ансамбля агентов, работающих на локально развернутой LLM Qwen-32B:q6 с различными параметрами генерации (top-p, top-k, температура и проч.). Генерация каждого примера сопровождалась кросс-валидацией: варианты, не совпадающие в 95 % случаев между агентами, отбраковывались. Дальнейшая процедура обучения модели ruT5-base проводилась в два этапа: предварительная общая оптимизация параметров на всем датасете и уточнение модели на контекстно-зависимых случаях. Такой подход обеспечивает предсказуемость и дополнительный контроль над процессом предобработки текста, что критично для образовательного контента. Однако его эффективность напрямую зависит от качества синтетических данных, что поднимает вопросы масштабируемости метода для узких доменов. Это требует дальнейшего расширения датасета за счет включения узкоспециализированных текстов и оптимизации баланса между обобщающей способностью модели и точностью воспроизведения экспертных правил. Во избежание частого дообучения модели такой способ разумно использовать в задачах с редкой сменой домена.

На выходе препроцессора генерируется окончательный текст лекции, к которому применяется модуль синтеза речи. На вход TTS-модуля поступает подготовленный текст и репрезентация голоса преподавателя, на выходе формируется синтезированная аудиодорожка с голосом, близким к оригинальному голосу преподавателя. Образец исходного голоса дополнительно подается в легковесную модель-модификатор звука для шумоподавления и нормализации. Затем из очищенного аудиофрагмента извлекаются эмбединги голоса (скрытые численные векторы) [16], которые служат голосовой репрезентацией преподавателя. Пользователь может предоставить один или несколько фрагментов голоса диктора с разными тембром и интонацией для выбора наиболее подходящего образца в дальнейшем синтезе. После этого полученные эмбединги передаются в TTS

для тонкой настройки тембра генерации синтезированного голоса под конкретного преподавателя. Для эффективной и удобной персонализации используется синтез zero-shot. При выборе архитектуры TTS учитывались следующие ключевые факторы: возможность zero-shot TTS-реализации и мультязычная поддержка, выражающаяся в способности корректно воспроизводить заимствованные термины из других языков и снижать объем требуемых данных для эффективного обучения.

Современные TTS-архитектуры делятся на специализированные (монолингвальные) и мультязычные. Специализированные модели требуют меньше данных и являются более надежными в рамках работы только с одним языком, однако плохо обрабатывают слова из других языков. Мультязычные модели могут воспроизводить фразы на разных языках, но они гораздо сложнее в обучении, и иногда появляются проблемы при переходе между языками. В рамках данной работы было решено использовать для произношения всего текста одну мультязычную модель, соответствующую архитектуре XTTS.

Фундаментальным ограничением всех современных TTS-систем является сложность обработки омографичных лексем. Различие между написанием слова и вероятным наличием у него множественных фонетических реализаций создает в процессе генерации неоднозначность, разрешение которой остается очень сложной задачей, непосредственно влияющей на точность и натуральность синтеза речи. Генеративный процесс носит вероятностный характер, поэтому зачастую отсутствие строгого гарантированного управления ударением в словах создает трудности. Для повышения качества произношения в сгенерированном аудио перед процессом синтеза с помощью TTS был использован модуль перевода текста в IPA с учетом ударений. Это позволило задать их явным образом: любые текстовые фразы переводятся в последовательность символов IPA с пометками ударения, после чего TTS-модель синтезирует речь на основе этой транскрипции. Таким образом, проблема омографов и ударений решается на уровне препроцессора, а синтезирующая модель обучается на унифицированном фонетическом представлении текста, полученном из специальной обработки данных из открытых датасетов (RUSLAN (<https://ruslan-corporus.github.io/>), CommonVoice (<https://commonvoice.mozilla.org/>) и SOVA – открытый набор размеченных аудиоданных для обучения

моделей синтеза речи (<https://sova.ai/ru/dataset/>) в формате LJ Speech.

В результате сравнений на тестовой выборке наиболее подходящим оказалось решение CoquiTTS (<https://coquitts.com/>), демонстрирующее приемлемое качество естественности и идентичности голоса при ограниченных вычислительных ресурсах и позволяющее гибко регулировать скорость генерируемой речи специальным параметром.

Генерация видеопортрета. Если от преподавателя не поступил готовый видеофайл образца чтения им обучающих материалов, система может автоматически создать реалистичный видеопортрет вместе с естественными движениями по одной фотографии, на которой отчетливо видно его лицо. В данной работе за генерацию отвечает открытый фреймворк LivePortrait. На вход подается портретная фотография преподавателя, а на выходе получается видеоролик с движениями головы. Фреймворк построен на модели ключевых точек лица и дает естественно выглядящий результат за время, сопоставимое с работой модуля ASR, что удовлетворяет всем потребностям, так как может вычисляться параллельно с этим модулем. Кроме того, реализация создания видео из фотографии может предоставлять гибкие генеративные возможности для контроля внешних стилистических атрибутов, таких как одежда и окружающая среда, за пределами реального образца преподавателя.

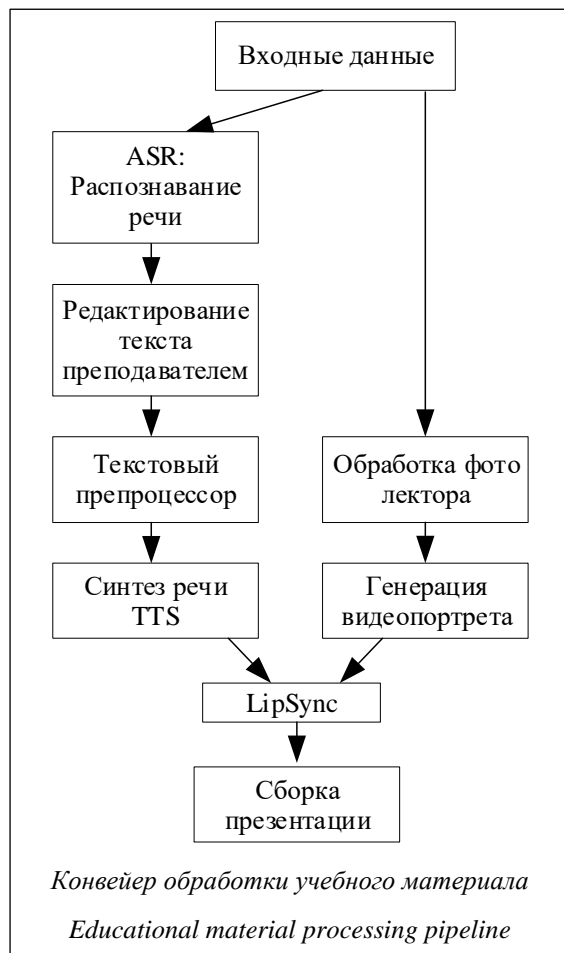
Синхронизация губ. После получения синтезированного голоса преподавателя с помощью TTS-системы речи и видеопортрета, представляющего облик преподавателя, оба этих потока данных подаются на вход LipSync-модуля. Этот модуль синхронизирует движения губ и лица цифрового аватара со звуковой последовательностью. Преимущество LatentSync модели заключается в устойчивости ко внешним TTS-источникам и в более высокой естественности артикуляции без необходимости постобработки. Однако, как и другие современные решения в этой области, она основана на покандровой генерации, что ограничивает производительность. В связи с этим предусмотрен режим адаптивной генерации: на первом этапе может применяться более быстрая, но менее точная модель, после чего при необходимости наиболее важные фрагменты повторно обрабатываются LatentSync для повышения визуального качества. Такой подход обеспечивает гибкость в зависимости от требований – от быстрой черновой генерации до высококачественного

рендеринга для финального образовательного продукта.

После генерации видеосегментов лекционного материала система автоматически интегрирует их в лекционный файл. Компонент сборки презентации выполняет операции вставки и извлечения внутри слайдов: к каждому слайду, где необходимо, прикрепляется соответствующий видеоролик с цифровым аватаром преподавателя. На выходе формируется модифицированная версия презентации, в которой слайды дополнены сгенерированными видео. Кроме того, по предпочтению пользователя можно сохранить итоговый образовательный материал, представленный целиком в видеоформате, с помощью экспорта всей презентации.

Структура линейного конвейера обработки данных. Предложенная система реализует упорядоченный последовательный конвейер обработки лекционного материала, состоящий из семи этапов, идущих друг за другом.

На рисунке представлена блок-схема, визуализирующая разработанный конвейер обработки данных. В результате его выполнения формируется обновленная презентация с инте-



грированными видеосегментами преподавателя и при необходимости экспортируется итоговый видеоролик полной лекции. На рисунке каждый блок соответствует отдельному модулю системы, продемонстрированы различные этапы обработки, которые могут выполняться параллельно.

Описание эксперимента

Для оценки качества работы каждого модуля и системы в целом были проведены вычислительные эксперименты. Для функционального тестирования была сформирована тестовая БД на основе реальных лекционных материалов, созданных для цифровой образовательной платформы Мирера [17, 18]. Эта база включала презентации с текстом и изображениями, фотографии и отрывки реальных записей голоса преподавателей для генерации видеопортретов.

При помощи данной базы проводилось тестирование следующих аспектов.

- **Скорость работы модулей:** измерялось время, требуемое для распознавания речи, а также затраченное на синтез речи zero-shot TTS-моделью, и время, необходимое для генерации видео лекционного материала. Результаты показали, что разработанная система существенно сокращает усилия преподавателей для создания и обновления учебных материалов по сравнению с традиционной перезаписью лекций, сохраняя при этом обобщенное качество на приемлемом уровне. Пользовательский сценарный анализ показал, что время, затраченное на пересоздание видео лекции с помощью описанных модулей, сократилось на ~ 70 % по сравнению с полной ее перезаписью.

- **Качество синтеза голоса:** аудиофрагменты оценивались по естественности и сходству с оригинальным голосом преподавателя. Разработанная и дообученная TTS-модель в совокупности с транскрипцией текста в IPA показали высокое качество сгенерированной речи в режиме zero-shot, обеспечивая в то же время высокую идентичность голоса при тонкой настройке скорости речи под характеристики диктора. Усредненная метрика естественности речи, по опросу, равна 4.12.

- **Качество видеогенерации и LipSync:** генерируемые видео с движениями лица оценивались по естественности выражения и синхронизации с аудио. Лучшая по LipScore модель LatentSync (метрика равна 4.21) продемонстрировала более точную синхронизацию, хотя

и требовала значительно большей вычислительной мощности. В целом эксперименты подтвердили, что конечный видеоряд воспринимается как достаточно естественный.

• Пользовательский опыт: преподаватели, знакомые ранее с цифровой образовательной платформой, отмечали удобство быстрой генерации обновленных лекций и сохранение узнаваемости своего голоса и образа в видео.

Экспериментальные результаты показали, что использование предложенной системы позволяет значительно сократить время подготовки учебных видео при сохранении приемлемого уровня естественности синтезированного аудио- и видеоряда. В процессе тестирования также были выявлены ограничения и точки роста: низкая эффективность обработки математических формул и специфических сокращений (что затрудняет генерацию речей из научных лекций), а также необходимость ускорения LipSync-модуля за счет более быстрой интерполяции между ключевыми кадрами.

Заключение

В рамках проведенного исследования предложена и реализована модульная архитектура персонального цифрового двойника преподавателя, включающая основные компоненты: ASR-модуль, универсальный текстовый препроцессор, TTS-модель, модуль генерации видеопортрета, LipSync и элемент сборки презентации. Особое внимание уделено модулю синтеза речи: требования zero-shot синтеза голоса нового диктора, мультиязычности и точности произношения на русском языке были обеспечены выбором открытой архитектуры XTTS в сочетании с IPA-предобработкой и инструментом расстановки ударений (<https://github.com/omogt/omogre>). Проведенное сравнение показало, что CoquiTTS обеспечивает приемлемое качество натуральности и идентичности голоса при ограниченных ресурсах, а также позволяет настраивать скорость речи для подстраивания под конкретного диктора. Интегрированный с TTS текстовый препроцессор уменьшил ошибки синтеза, связанные с омографами, числительными и сокращениями, что критично для образовательного контента.

Исследование подтвердило необходимость разработки собственного препроцессора для русскоязычных TTS-систем, учитывающего ударения, числительные и сокращения. Сравнение двух подходов – дообучения компактной модели ruT5-base и использования крупной

LLM с промпт-инжинирингом – показало их применимость в разных условиях с предобработкой в различных доменах. Модель RuT5 демонстрирует высокую точность в статичных доменах, но требует значительных усилий по генерации и обновлению данных. В работе выбор сделан в пользу применения Qwen-32B:q6, гарантирующей гибкость и быструю адаптацию к динамически меняющимся требованиям (например, при переходе между гуманитарными и техническими дисциплинами). Такой подход обеспечивает оптимальный баланс мультиязычности, скорости внедрения и совместимости с существующими LLM-технологиями. Модификация системных промптов позволяет оперативно добавлять правила для специализированных конструкций (например, для математических формул или технических аббревиатур) без переобучения модели. Для узкоспециализированных сценариев предложен гибридный режим: LLM выполняет первичную обработку, а дообученная модель ruT5-base уточняет специфичные конструкции. В дальнейшем планируется автоматизировать выбор между этими подходами на основе анализа семантики текста и оптимизировать их обучение. Выбор CoquiTTS и IPA-подхода подтвердил, что унификация входного текста через фонетическую транскрипцию позволяет переложить проблемы мультиязычности и ударений на задачу качественного препроцессора.

Предложенное решение вносит вклад как в прикладную область автоматизации создания образовательных видео, так и в исследования TTS для русскоязычных задач. В работе описаны методы синтетической генерации данных для дообучения фонетической модели, обоснован выбор zero-shot архитектуры, разработаны практические рекомендации по интеграции IPA-предобработки и гибридного взаимодействия LLM с компактными моделями для текстовой подготовки. Рассмотрена перспективность применения модульных генеративных систем в образовательных платформах, что может способствовать дальнейшему развитию дистанционных образовательных технологий.

Развитие системы возможно в следующих направлениях:

- разработка специализированных алгоритмов распознавания и фонетической нормализации формул и сокращений перед синтезом речи;
- дообучение препроцессора и TTS-модуля на корпусах с терминологией иностранных языков для поддержки лекций в смешанных лингвистических средах;

– аннотация и добавление в тренировочные датасеты редких слов и конструкций, а также внедрение механизмов постфильтрации для повышения корректности произношения.

Кроме того, остаются открытыми для решения такие задачи, как автоматизация динамического переключения между LLM и ruT5-base на

основе семантического анализа входного текста для оптимального баланса скорости и точности, а также исследование методов параллельной или приближенной генерации мимики (например, через интерполяцию ключевых кадров) для ускорения конвейера без значительной потери качества.

Список литературы

1. Nassif A.B., Shahin I., Attili I., Azzeh M. et al. Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 2019, vol. 7, pp. 19143–19165. doi: 10.1109/access.2019.2896880.
2. Devlin J., Chang M.-W., Lee K., Toutanova K. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc. Conf. NAACL*, 2018, vol. 1, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
3. Lee M., Chang J.-H., Lee S.-E., Seong J.-S. et al. Regularizing Transformer-Based acoustic models by penalizing attention weights. *Proc. Interspeech*, 2022, pp. 56–60. doi: 10.21437/interspeech.2022-362.
4. Hrinchuk O., Popova M., Ginsburg B. Correction of automatic speech recognition with transformer sequence-to-sequence model. *IEEE ICASSP*, 2020, pp. 7074–7078. doi: 10.1109/ICASSP40776.2020.9053051.
5. Baade A., Peng P., Harwath D. MAE-AST: Masked autoencoding audio spectrogram transformer. *Proc. Interspeech*, 2022, pp. 2438–2442. doi: 10.21437/Interspeech.2022-10961.
6. Xie Y., Xu H., Song G. et al. X-Portrait: Expressive portrait animation with hierarchical motion attention. *Proc. SIGGRAPH Conf. Papers*, 2024, art. 115. doi: 10.1145/3641519.3657459.
7. Wang T.-C., Mallya A., Liu M.-Y. One-shot free-view neural talking-head synthesis for video conferencing. *Proc. IEEE/CVF Conf. CVPR*, 2021, pp. 10034–10044. doi: 10.1109/CVPR46437.2021.00991.
8. Tokuda K., Nankaku Y., Toda T. et al. Speech synthesis based on hidden Markov models. *Proc. of the IEEE*, 2013, vol. 101, no. 5, pp. 1234–1252. doi: 10.1109/jproc.2013.2251852.
9. Ping W., Peng K., Gibiansky A. et al. Deep Voice 3: Scaling text-to-speech with convolutional sequence learning. *ArXiv*, 2018, art. 1710.07654. URL: <https://arxiv.org/abs/1710.07654> (дата обращения: 05.06.2025).
10. Chen S., Wang C., Wu Y. et al. Neural codec language models are zero-shot text to speech synthesizers. *IEEE TASLPRO*, 2025, vol. 33, pp. 705–718. doi: 10.1109/TASLPRO.2025.3530270.
11. Casanova E., Davis K., Gulea I. et al. XTTS: A massively multilingual zero-shot text-to-speech model. *Proc. Interspeech*, 2024, pp. 4978–4982. doi: 10.21437/Interspeech.2024-2016.
12. Prajwal K.R., Mukhopadhyay R., Nambodiri V.P., Jawahar C.V. A Lip Sync expert is all you need for speech to lip generation in the wild. *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 484–492. doi: 10.1145/3394171.3413532.
13. Wang X., Li Y., Zhang H., Shan Y. Towards real-world blind face restoration with generative facial prior. *Proc. IEEE/CVF CVPR*, 2021, pp. 9164–9174. doi: 10.1109/CVPR46437.2021.00905.
14. Bigata A., Mira R., Bounareli S. et al. Keyface: Expressive audio-driven facial animation for long sequences via KeyFrame interpolation. *ArXiv*, 2025, art. 2503.01715v2. URL: <https://arxiv.org/html/2503.01715v2> (дата обращения: 05.06.2025).
15. Zmitrovich D., Abramov A., Kalmykov A. et al. A family of pretrained transformer language models for Russian. *ArXiv*, 2024, art. 2309.10931v4. URL: <https://arxiv.org/abs/2309.10931v4> (дата обращения: 05.06.2025).
16. Snyder D., Ghahremani P., Povey D. et al. Deep neural network-based speaker embeddings for end-to-end speaker verification. *Proc. IEEE SLT*, 2016, pp. 165–170. doi: 10.1109/SLT.2016.7846260.
17. Васильев И.А., Кушниренко А.Г., Леонов А.Г. и др. Цифровая образовательная платформа Мирера – основа цифровой трансформации образовательного процесса // Новые образовательные стратегии в современном информационном пространстве: сб. науч. статей по матер. Междунар. науч.-практич. конф. 2023. С. 140–144.
18. Дьяченко М.С., Леонов А.Г. Решение задачи автоматизации учебного процесса с помощью экспериментального поиска индивидуальной образовательной траектории // Информатика и образование. 2024. Т. 39. № 4. С. 14–26. doi: 10.32517/0234-0453-2024-39-4-14-26.

Automated generation of educational video materials based on a modular digital twin of the instructor

Aleksandr G. Leonov^{1, 2, 3, 4✉}, Kirill A. Mashchenko^{1, 2, 3}, Nikolay S. Martynov²,
Konstantin K. Pchelin², Artem V. Shlyakhov²

¹ Lomonosov Moscow State University, Moscow, 119991, Russian Federation

² National Research Centre “Kurchatov Institute”, SRISA, Moscow, 117218, Russian Federation

³ Moscow State University of Education, Moscow, 119991, Russian Federation

⁴ State University of Management, Moscow, 109542, Russian Federation

For citation

Leonov, A.G., Mashchenko, K.A., Martynov, N.S., Pchelin, K.K., Shlyakhov, A.V. (2025) ‘Automated generation of educational video materials based on a modular digital twin of the instructor’, *Software & Systems*, 38(4), pp. 694–703 (in Russ.). doi: 10.15827/0236-235X.152.694-703

Article info

Received: 20.06.2025

After revision: 15.07.2025

Accepted: 24.07.2025

Abstract. The paper presents a system for a personal digital twin of a university instructor, designed for the automated creation and updating of educational video materials within a digital learning platform. The proposed system architecture consists of a set of interconnected modules, each addressing a specific task and ensuring both flexibility and scalability. The first stage implements automatic speech recognition using the open-source multilingual Whisper model. The system handles original audio recordings without requiring prior domain-specific fine-tuning. The system then processes the retrieved text through a universal preprocessor, which performs three key functions: numeral normalization, word stress prediction, and transcription into the International Phonetic Alphabet (IPA). Speech synthesis is implemented using a modified CoquiTTS architecture supporting zero-shot voice cloning from a short instructor speech sample. The LivePortrait technology generates the visual component by creating realistic video portraits of the instructor using submitted photographs. The LipSync module synchronizes lip movements with generated audio, significantly improving perceived video quality. The final stage involves the automatic assembly of a presentation that integrates the generated video segments. The paper presents a comparative analysis of alternative architectures for each module, identifies evaluation criteria for speech synthesis and video generation quality, and outlines the specifics of developing a Russian-language text preprocessor. Experimental results demonstrate that the proposed system reduces instructor workload for the creation and updating of educational materials by more than 70% while maintaining a high level of naturalness. The authors discussed limitations in text processing of formulas and abbreviations, while also examining potential optimizations for improving the performance of the LipSync module.

Keywords: digital twin, text preprocessor, educational video materials, automatic speech recognition (ASR), text-to-speech (TTS), LipSync, zero-shot voice cloning, Russian-language TTS

Acknowledgements. The work was carried out within the framework of the state assignment National Research Centre “Kurchatov Institute” – SRISA on the topic no. FNEF-2024-0001 (1023032100070-3-1.2.1)

References

1. Nassif, A.B., Shahin, I., Attili, I., Azzeh, M. et al. (2019) ‘Speech recognition using deep neural networks: A systematic review’, *IEEE Access*, 7, pp. 19143–19165. doi: 10.1109/access.2019.2896880.
2. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. et al. (2018) ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, *Proc. Conf. NAACL*, 1, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
3. Lee, M., Chang, J.-H., Lee, S.-E., Seong, J.-S. et al. (2022) ‘Regularizing Transformer-Based acoustic models by penalizing attention weights’, *Proc. Interspeech*, pp. 56–60. doi: 10.21437/interspeech.2022-362.
4. Hrinchuk, O., Popova, M., Ginsburg, B. (2020) ‘Correction of automatic speech recognition with transformer sequence-to-sequence model’, *IEEE ICASSP*, pp. 7074–7078. doi: 10.1109/ICASSP40776.2020.9053051.
5. Baade, A., Peng, P., Harwath, D. (2022) ‘MAE-AST: Masked autoencoding audio spectrogram transformer’, *Proc. Interspeech*, pp. 2438–2442. doi: 10.21437/Interspeech.2022-10961.
6. Xie, Y., Xu, H., Song, G. et al. (2024) ‘X-Portrait: Expressive portrait animation with hierarchical motion attention’, *Proc. SIGGRAPH Conf. Papers*, art. 115. doi: 10.1145/3641519.3657459.
7. Wang, T.-C., Mallya, A., Liu, M.-Y. (2021) ‘One-shot free-view neural talking-head synthesis for video conferencing’, *Proc. IEEE/CVF Conf. CVPR*, pp. 10034–10044. doi: 10.1109/CVPR46437.2021.00991.
8. Tokuda, K., Nankaku, Y., Toda, T. et al. (2013) ‘Speech synthesis based on hidden Markov models’, *Proc. of the IEEE*, 101(5), pp. 1234–1252. doi: 10.1109/jproc.2013.2251852.
9. Ping, W., Peng, K., Gibiansky, A. et al. (2018) ‘Deep Voice 3: Scaling text-to-speech with convolutional sequence learning’, *ArXiv*, art. 1710.07654, available at: <https://arxiv.org/abs/1710.07654> (accessed June 5, 2025).
10. Chen, S., Wang, C., Wu, Y. et al. (2025) ‘Neural codec language models are zero-shot text to speech synthesizers’, *IEEE TASLPRO*, 33, pp. 705–718. doi: 10.1109/TASLPRO.2025.3530270.
11. Casanova, E., Davis, K., Gulea, I. et al. (2024) ‘XTTS: A massively multilingual zero-shot Text-to-Speech model’, *Proc. Interspeech*, pp. 4978–4982. doi: 10.21437/Interspeech.2024-2016.
12. Prajwal, K.R., Mukhopadhyay, R., Nambodiri, V.P., Jawahar, C.V. (2020) ‘A Lip Sync expert is all you need for speech to lip generation in the wild’, *ACM Int. Conf. Multimedia*, pp. 484–492. doi: 10.1145/3394171.3413532.
13. Wang, X., Li, Y., Zhang, H., Shan, Y. (2021) ‘Towards real-world blind face restoration with generative facial prior’, *Proc. IEEE/CVF CVPR*, pp. 9164–9174. doi: 10.1109/CVPR46437.2021.00905.
14. Bigata, A., Mira, R., Bounareli, S. et al. (2025) ‘Keyface: Expressive audio-driven facial animation for long sequences via KeyFrame interpolation’, *ArXiv*, art. 2503.01715v2, available at: <https://arxiv.org/html/2503.01715v2> (accessed June 5, 2025).
15. Zmitrovich, D., Abramov, A., Kalmykov, A. et al. (2024) ‘A family of pretrained transformer language models for Russian’, *ArXiv*, art. 2309.10931v4, available at: <https://arxiv.org/abs/2309.10931v4> (accessed June 5, 2025).
16. Snyder, D., Ghahremani, P., Povey, D. et al. (2016) ‘Deep neural network-based speaker embeddings for end-to-end speaker verification’, *Proc. IEEE SLT*, pp. 165–170. doi: 10.1109/SLT.2016.7846260.

17. Vasilyev, I.A., Kushnirenko, A.G., Leonov, A.G. et al. (2023). 'Mirera digital educational platform - the basis of digital transformation of the educational process', *Proc. Sci.-Pract. Conf. New Educational Strategies in Modern Inform. Space*, pp. 140–144 (in Russ.).

18. Diachenko, M.S., Leonov, A.G. (2024) 'Solving the problem of automating the learning process through experimental search for an individual educational trajectory', *Informatics and Education*, 39(4), pp. 14–26 (in Russ.). doi: 10.32517/0234-0453-2024-39-4-14-26.

Авторы

Леонов Александр Георгиевич^{1, 2, 3, 4},

д.п.н., к.ф.-м.н., профессор,
доцент, зав. кафедрой,
ведущий научный сотрудник,
dr.l@math.msu.ru

Машенко Кирилл Алексеевич^{1, 2, 3},

младший научный сотрудник,
kirill.mashchenko@niisi.ru

Мартынов Николай Сергеевич²,

младший научный сотрудник,
nikolai.martynov@math.msu.ru

Пчелин Константин Константинович²,

техник, konstantin.pchelin@math.msu.ru

Шляхов Артем Вячеславович²,

младший научный сотрудник,
shlyakhov@vip.niisi.ru

Authors

Aleksandr G. Leonov^{1, 2, 3, 4}, Dr.Sci. (Education),

Cand. of Sci. (Physics and Mathematics),
Professor, Associate Professor,
Head of Chair, Leading Researcher,
dr.l@math.msu.ru

Kirill A. Mashchenko^{1, 2, 3},

Junior Researcher,
kirill.mashchenko@niisi.ru

Nikolay S. Martynov²,

Junior Researcher,
nikolai.martynov@math.msu.ru

Konstantin K. Pchelin², Technical Fellow,

konstantin.pchelin@math.msu.ru

Artem V. Shlyakhov²,

Junior Researcher,
shlyakhov@vip.niisi.ru

¹ МГУ имени М.В. Ломоносова,
г. Москва, 119991, Россия

² НИЦ «Курчатовский институт» – НИИСИ,
г. Москва, 117218, Россия

³ Московский педагогический государственный
университет, г. Москва, 119991, Россия

⁴ Государственный университет управления,
г. Москва, 109542, Россия

¹ Lomonosov Moscow State University,
Moscow, 119991, Russian Federation

² National Research Centre "Kurchatov Institute" – SRISA,
Moscow, 117218, Russian Federation

³ Moscow State University of Education,
Moscow, 119991, Russian Federation

⁴ State University of Management,
Moscow, 109542, Russian Federation